SVEUČILIŠTE U ZAGREBU

PRIRODOSLOVNO–MATEMATIČKI FAKULTET

MATEMATIČKI ODSJEK

Marina Bišćan

PREPOZNAVANJE I ANALIZA SPOLA PREMA AKUSTIČNIM KARAKTERISTIKAMA GLASA I GOVORA

Diplomski rad

Voditelj rada:
prof. dr. sc. Anamarija Jazbec

Zagreb, rujan, 2017

Ovaj diplomski rad obranjen je dana _____ pred ispitnim povjerenstvom u sastavu:

1. _____, predsjednik
2. _____, član
3. _____, član

Povjerenstvo je rad ocijenilo ocjenom _____.

Potpisi članova povjerenstva:

1. _____
2. _____
3. _____

Posebno se želim zahvaliti svojoj mentorici prof. dr. sc. Anamariji Jazbec na pristupačnosti i savjetima koje mi je pružila prilikom izrade ovog diplomskog rada. Puno se zahvaljujem svojoj obitelji na pruženoj ljubavi i potpori i na svemu što su mi pružili. Veliko hvala mojem dečku koji mi je bio velika pomoć i podrška kroz sve godine studiranja. Također se želim puno zahvaliti svojim prijateljima što su vjerovali u mene i bili uz mene u svim trenucima.

Sadržaj

Sadržaj **iv**

Uvod **2**

1 Logistička regresija **3**
1.1 Regresijska analiza 3
1.2 Linearna regresija 3
1.3 Povezanost linearne i logističke regresije 5
1.4 Šansa (eng. *Odds*) i omjer šansi (eng. *Odds Ratio*) 8
1.5 Logit model 10
1.6 Procjena parametara u modelu logističke regresije 11
1.7 Testiranje adekvatnosti modela (eng. *Goodness of fit*) 14
1.8 *ROC* krivulja 14

2 Klasifikacijska i regresijska stabla odlučivanja **16**
2.1 Općenito 16
2.2 Klasifikacijsko stablo 17

3 Prepoznavanje i analiza spola prema akustičnim karakteristikama glasa i govora **27**
3.1 Pojmovnik 27
3.2 Opis podataka 29
3.3 Desktiptivna statistika 30
3.4 Univarijatna logistička regresija 43
3.5 Multivarijatna logistička regresija 57
3.6 Primjena CART analize 63

Bibliografija **71**

Uvod

Određivanje spola osobe na temelju uzorka njihovog glasa se čini kao jednostavan zadatak. Često, ljudsko uho može lako otkriti razliku između muškog i ženskog glasa nakon svega nekoliko izgovorenih riječi. S druge strane, napraviti kompjuterski program koji može odrediti spol osobe prema njezinom glasu je već malo teže. Akustična analiza glasa ovisi o svojstvima parametara specifičnih za uzorak poput intenziteta, trajanja, frekvencije i sličnih svojstava. Akustična svojstva glasa i govora mogu se koristiti za određivanje spola osobe. Programski jezik R sadrži pakete "warbleR", "seewave" i "tuneR" koji su napravljeni za akustičnu analizu. Koristeći navedene analize, dobili smo bazu podataka koja će se koristiti u ovom radu.

U ovom radu pomoću logističke regresije analizirati ćemo akustične karakteristike prema spolu i procijeniti koje od prikupljenih varijabli statistički značajno utječu na prepoznavanje spola. Također ćemo pomoću klasifikacijskih stabala analizirati akustične karakteristike glasa i govora te pokušati potvrditi rezultate koje smo dobili koristeći logističku regresiju.

Logističku regresiju razvio je statističar David Cox 1958. godine. Binarni logistički model koristi se za procjenu vjerojatnosti binarnog odgovora u odnosu na jednu ili više prediktornih (ili nezavisnih) varijabli. Može se reći da prisutnost faktora rizika povećava vjerojatnost danog ishoda određenim postotkom.
Stablo odlučivanja je metoda koja se obično koristi u rudarenju podataka (eng. *data mining*). Cilj je stvoriti model koji predviđa vrijednost ciljane varijable na temelju nekoliko ulaznih varijabli. Stabla odlučivanja koja se koriste u rudarenju podataka su: analiza klasifikacijskog stabla koja se koristi kada je predviđeni ishod kategorije kojoj pripada podatak, te analiza regresijskog stabla koja se koristi kada se predviđeni ishod može smatrati stvarnim brojem. Statističar L. Breiman se smatra izumiteljem CART (eng. *Classification and Regression Tree*) analize.

U prvom poglavlju je opisana teorijska pozadina modela logističke regresije. Prvo je ukratko opisana regresijska analiza, linearna regresija i logistička regresija gdje

ćemo definirati osnovne pojmove te ćemo objasniti metodu maksimalne vjerodostojnosti, omjer šanse i testiranje adekvatnosti modela.
U drugom poglavlju uvodimo pojam analize pomoću klasifikacijskog i regresijskog stabla odlučivanja, te detaljnije opisujemo teorijsku pozadinu klasifikacijskog stabla.
U trećem poglavlju koristeći programski jezik SAS opisanu teorijsku pozadinu modela iz prvog i drugog poglavlja primjenjujemo na bazu podataka koja sadrži akustične karakteristike glasa i govora kako bismo analizirali spol.

Poglavlje 1

Logistička regresija

1.1 Regresijska analiza

U statističkom modeliranju, regresijska analiza je statistički proces za procjenu veza između varijabli. To je metoda ispitivanja i analize ovisnosti jedne ili više zavisnih varijabli o jednoj ili više drugih (nezavisnih) varijabli. Jedan od rezultata svake regresijske analize je regresijski model. Regresijski model je matematička jednadžba koja definira tj. kvantificira povezanost između zavisne varijable s nezavisnim. Ako je povezanost između zavisne i nezavisne varijable linearna govorimo o linearnoj regresiji. Regresija ne mora biti linearna u tom slučaju govorimo o nelinearnoj regresiji.

Neka je Y zavisna varijabla (varijabla odaziva) koju želimo procijeniti ili opisati, te neka je X nezavisna varijabla (varijabla poticaja ili prediktorna varijabla) pomoću koje želimo opisati zavisnu varijablu, tada govorimo o univarijatnoj (jednostrukoj) regresijskoj analizi. Ukoliko ispitujemo ovisnost zavisne varijable o n nezavisnih varijabli, pri čemu je $n \geq 2$, onda govorimo o multivarijatnoj (višestrukoj) regresijskog analizi.

Regresijska analiza u kojoj je zavisna varijabla diskretna (dihotomna) tj. može poprimiti dvije ili više vrijednosti zovemo logistička regresija ili logit model. Kod linearne regresije povezanost između jedne zavisne i jedne nezavisne varijable opisana je jednadžbom pravca, dok je povezanost između jedne zavisne i više nezavisnih varijabli opisana jednadžbom ravnine.

1.2 Linearna regresija

Neka su $x^{(1)}, x^{(2)}, \ldots, x^{(k)}$ kontrolirane (neslučajne) varijable i Y slučajna varijabla mjerena u ovisnosti o $x = (x^{(1)}, x^{(2)}, \ldots, x^{(k)})$, odnosno $Y = Y(x)$. Linearni model

ovisnosti veličine Y o x zadan je sa

$$Y = \beta_0 + \beta_1 x^{(1)} + \cdots + \beta_k x^{(k)} + \varepsilon, \tag{1.1}$$

gdje je ε slučajna pogreška, a $\beta_0, \beta_1, \ldots, \beta_k$ parameteri modela. Općeniti zapis je

$$Y = \beta_0 + \beta_1 p_1(x) + \cdots + \beta_k p_k(x) + \varepsilon, \tag{1.2}$$

gdje su $1, p_1, \ldots, p_k$ linearno nezavisne realne funkcije. Stavimo sljedeće oznake, $x_i = (x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(k)})$ za $i = 1, 2, \ldots, n$ zadane vrijednosti od x takve da su barem dvije različite i $y_1, y_2, \ldots, y_n$ realizacije slučajne varijable za Y. Minimiziramo funkciju

$$L(\beta_0, \beta_1, \ldots, \beta_k) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i^{(1)} - \cdots - \beta_k x_i^{(k)})^2, \tag{1.3}$$

odnosno općenitu funkciju

$$L(\beta_0, \beta_1, \ldots, \beta_k) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 p_1(x_i) \cdots - \beta_k p_k(x_i))^2. \tag{1.4}$$

Neka su $Y_i = \beta_0 + \beta_1 x_i^{(1)} + \cdots + \beta_k x_i^{(k)} + \varepsilon_i$, za $i = 1, 2, \ldots, n$, slučajne varijable. Pretpostavimo da vrijede Gauss-Markovljevi uvjeti:

- $\mathbb{E}[\varepsilon_i] = 0, \forall i = 1, 2, \ldots, n$
- $\mathbb{E}[\varepsilon_i \varepsilon_j] = 0, \forall i, j = 1, 2, \ldots, n$ takve da je $i \neq j$ (*nekoreliranost*)
- $Var(\varepsilon_i) = \sigma^2 > 0, \forall i = 1, 2, \ldots, n$

Stavimo:

$$X = \begin{bmatrix} 1 & x_1^{(1)} & x_1^{(2)} & \ldots & x_1^{(k)} \\ 1 & x_2^{(1)} & x_2^{(2)} & \ldots & x_2^{(k)} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_n^{(1)} & x_n^{(2)} & \ldots & x_n^{(k)} \end{bmatrix}$$

ili općenitije

$$X = \begin{bmatrix} 1 & p_1(x_1) & p_2(x_1) & \ldots & p_k(x_1) \\ 1 & p_1(x_2) & p_2(x_2) & \ldots & p_k(x_2) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & p_1(x_n) & p_2(x_n) & \ldots & p_k(x_n) \end{bmatrix}$$

i

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}, \quad y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_k \end{bmatrix}, \quad Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_k \end{bmatrix}, \quad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_k \end{bmatrix}.$$

Tada je

$$L(\beta) = ||y - X\beta||^2. \tag{1.5}$$

Nepristrani procjenitelj za β metodom najmanjih kvadrata je

$$\hat{\beta} = (X^T X)^{-1} X^T Y \tag{1.6}$$

uz procjenu $(X^T X)^{-1} X^T y$, a procjenitelji za Y_i su

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i^{(1)} + \cdots + \hat{\beta}_k x_i^{(k)}, \text{ za } i = 1, 2, \ldots, n, \tag{1.7}$$

odnosno općenitije

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 p_1(x_i) + \cdots + \hat{\beta}_k p_k(x_i), \text{ za } i = 1, 2, \ldots, n. \tag{1.8}$$

Reziduali su slučajne varijable $E_i = Y_i - \hat{Y}_i$, odnosno njihove realizacije e_i, za $i = 1, 2, \ldots, n$. Dodatno pretpostavljamo da je $\varepsilon \sim Normal$. [7]

1.3 Povezanost linearne i logističke regresije

Pretpostavke na linearni model imamo u prethodnom poglavlju, ali navest ćemo ih još jednom radi boljeg pregleda:

1. $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$
2. $\mathbb{E}[\varepsilon_i] = 0, \forall i = 1, 2, \ldots, n$
3. $\mathbb{E}[\varepsilon_i \varepsilon_j] = 0, \forall i, j = 1, 2, \ldots, n$ takve da je $i \neq j$ (*nekoreliranost*)
4. $Var(\varepsilon_i) = \sigma^2 > 0, \forall i = 1, 2, \ldots, n$
5. $\varepsilon_i \sim Normal$

Kako bi nam bilo lakše, pretpostavimo da imamo jednu nezavisnu varijablu x i uz to pretpostavimo da je x fiksan kroz cijeli ponovljeni uzorak (što znači da svaki uzorak ima isti skup vrijednosti x-eva). Indeksom i razlikujemo različite članove uzorka. Pretpostavka 1 kaže da je y linearna funkcija od x plus slučajna pogreška

ε, za sve članove uzorka. Ostale pretpostavke govore nešto o pogreški ε, na primjer pretpostavka 2 implicira da su x i ε nekorelirane. Pretpostavka 4 se često zove *homoskedastičnost* i kaže da je varijanca od ε ista za sve observacije, a dok nam pretpostavka 3 kaže da slučajna pogreška jedne observacije je nekorelirana sa slučajnom pogreškom bilo koje druge observacije. Ukoliko su zadovoljene sve pretpostavke, metodom najmanjih kvadrata dobivamo da su β_0 i β_1 nepristrani procijenitelji i imaju minimalnu uzoračku varijancu (minimalna varijabilnost kroz ponovljeni uzorak).

Pretpostavimo sada da je y dihotomna varijabla, odnosno da može poprimiti vrijednosti 1 ili 0. I dalje je razumno tvrditi da su pretpostavke 1, 2 i 4 točne. Ukoliko su pretpostavke 1 i 2 točne za dihotomnu varijablu, onda su pretpostavke 3 i 5 nužno netočne. Prvo, pretpostavimo da vrijedi pretpostavka 5, te pretpostavimo da je $y_i = 1$. Tada pretpostavka 1 povlači $\varepsilon_i = 1 - \beta_0 - \beta_1 x_i$. S druge strane, ako je $y_i = 0$ onda imamo $\varepsilon_i = -\beta_0 - \beta_1 x_i$. Zbog toga jer ε_i možemo poprimiti samo dvije vrijednosti, nemoguće je da bude normalno distribuirano (što znači da je neprekidno i nema donju i gornju granicu). Stoga pretpostavka 5 mora biti odbačena. Kako bismo procijenili pretpostavku 3, korisno je napraviti malo osnovne algebre. Očekivanje od y_i je po definiciji

$$\mathbb{E}[y_i] = 1 \times \mathbb{P}(y_i = 1) + 0 \times \mathbb{P}(y_i = 0).$$

Ako definiramo $p_i = \mathbb{P}(y_i = 1)$, onda dobivamo

$$\mathbb{E}[y_i] = p_i. \tag{1.9}$$

Također pretpostavke 1 i 2 impliciraju sljedeći zapis za očekivanje.

$$\begin{aligned} \mathbb{E}[y_i] &= \mathbb{E}[\beta_0 + \beta_1 x_i + \varepsilon_i] \\ &= \mathbb{E}[\beta_0] + \mathbb{E}[\beta_1 x_i] + \mathbb{E}[\varepsilon_i] \\ &= \beta_0 + \beta_1 x_i \end{aligned} \tag{1.10}$$

Iz 1.9 i 1.10 dobivamo

$$p_i = \beta_0 + \beta_1 x_i \tag{1.11}$$

i ovaj rezultat se ponekad zove *linearni vjerojatnosni model*. Kako ime sugerira, ovaj model sugerira da je vjerojatnost da je $y = 1$ linearna funkcija od x. Regresijski koeficijenti imaju jasnu interpretaciju ovog modela.
Sada razmotrimo varijancu od ε_i. Zato jer je x fiksan, varijanca od ε_i je ista kao i varijanca od y_i. Općenito, varijanca dihotomne varijable je $p_i(1 - p_i)$. Stoga, imamo

$$Var(\varepsilon_i) = p_i(1 - p_i) = (\beta_0 + \beta_1 x_i)(1 - \beta_0 - \beta_1 x_i). \tag{1.12}$$

Vidimo da varijanca od ε_i mora biti različita za različite observacije i ona varira kao funkcija x. Varijanca pogreške poprima maksimum kada je $p_i = 0.5$ i smanjuje se što je p_i bliži 1 ili 0. Upravo smo pokazali da dihotomna nezavisna varijabla u linearnom regresijskom modelu nužno ne zadovoljava pretpostavke za homoskedastičnost (pretpostavka 4) i normalnost (pretpostavka 5). Bitno pitanje koje se pitamo, koje su posljedice? Nisu toliko ozbiljne koliko mislimo. Prije svega, ove pretpostavke nam nisu potrebne da bismo dobili nepristrane procijenitelje. Ukoliko vrijede samo pretpostavke 1 i 2, metoda najmanjih kvadrata će nam dati nepristrane procijenitelje za β_0 i β_1. Drugo, pretpostavka normalnosti nam nije potrebna ukoliko je uzorak dovoljno veliki. Centralni granični teorem nam osigurava da procijenjeni koeficijenti imaju distribuciju koja je aproksimativno normalna onda kada ε nije normalno distribuiran. To upravo znači da možemo koristiti tablicu normalne distribucije kako bismo izračunali p-vrijednosti i pouzdane intervale. Ako je uzorak malen, ove aproksimacije bi mogle biti loše. Kada nije zadovoljena pretpostavka homoskedastičnosti imamo dvije nepoželjne posljedice. Prvo, procijenjeni koeficijenti više nisu efikasni. U statističkoj terminologiji, to znači da postoje alternativne metoda za procjenu sa manjim standardnim pogreškama. Drugo, i bitnije, procjenitelji standardne pogreške (eng. *standard error estimates*) nisu više konzistentni procjenitelji za točne standardne pogreške (eng. *true standard errors*). To znači da procjenitelji standardne pogreške mogu biti pristrani (bilo prema gore ili prema dolje) do nepoznatnog stupnja. Zbog toga jer se standardne pogreške koriste u računanju test statistika, test statistike bi također mogle biti pristrane.

Uz ove tehničke poteškoće, postoji fundamentalniji problem s pretpostavkama za linearni model. Vidjeli smo da ako imamo dihotomnu varijablu, onda pretpostavke 1 i 2 povlače linearni vjerojatnosni model $p_i = \beta_0 + \beta_1 x_i$. Zapravo s ovim modelom je sve u redu, ali je malo nemoguć pogotovo kada se x mjeri na neprekidnom skupu. Ako x nema gornju ili donju granicu, tada za bilo koju vrijednost od β_1 postoje vrijednosti od x za koje je p_i veći ili jednak od 1 ili manji od 0. U stvari, kada procijenjujemo linearni vjerojatnosti model sa metodom najmanjih kvadrata, česta je pojava da predviđene vrijednosti dobivene generiranjem budu izvan intervala $(0, 1)$. Naravno, nemoguće je za točne vrijednosti (koje su vjerojatnosti) budu veće ili jednake od 1 ili manje od 0. Jedini način da model bude točan je ako se uspije ograničiti p_i s gornje i donje strane, ali takve stvari stvaraju probleme teorijski i računski.

Statističari su zbog ovih problema s linearnim modelom razvili alternativne pristupe koji su konceptualno imali više smisla, te također bolja statistička svojstva. Najpopularniji pristup je *logit model* koji je procijenjen sa maksimalnom vjerodostojnosti. Prije nego što razmotrimo puni model, ispitat ćemo jednu od njegovih komponenti - šanse jednog događaja.

1.4 Šansa (eng. *Odds*) i omjer šansi (eng. *Odds Ratio*)

Da bismo cijenili logit model, korisno je razumijeti šanse i omjer šansi. Postoji jednostavna veza između vjerojatnosti i šanse. Ako s p označimo vjerojatnost da se neki događaj dogodio, a s O (eng.*Odds*) šansu tog događaja, tada imamo

$$\begin{aligned} O &= \frac{p}{1-p} = \frac{\text{vjerojatnost da se događaj dogodio}}{\text{vjerojatnost da se događaj nije dogodio}} \\ p &= \frac{O}{1+O} \end{aligned} \tag{1.13}$$

Odnos između vjerojatnosti i šanse je prikazan u tablici 1.1.

Tablica 1.1: Odnos između vjerojatnosti i šansi

Vjerojatnost	Šansa
0.10	0.11
0.20	0.25
0.30	0.43
0.40	0.67
0.50	1.00
0.60	1.50
0.70	2.33
0.80	4.00
0.90	9.00

Iz tablice 1.1 možemo vidjeti da šansa manja od 1 odgovara vjerojatnosti manjoj od 0.5, dok šanse veće od 1 odgovaraju vjerojatnosti većoj od 0.5. Kao i vjerojatnost, šanse imaju donju granicu 0, ali nemaju gornju granicu što je drugačije nego vjerojatnost. Kako bismo što bolje shvatili pojmove šanse i omjer šansi upotrijebit ćemo primjer koji je prikazan u tablici 1.2.

Tablica 1.2: Izgled tablice kontingencije rizičnog faktora i stanja bolesti

		Stanje		Ukupno
		Bolestan	Zdrav	
Rizičan faktor	Pije alkohol	a	b	a + b
	Ne pije alkohol	c	d	c + d
Ukupno		a + c	b + d	a + b + c + d

U primjeru s p označimo vjerojatnost da imamo neku bolest, a šansa da dobijemo tu bolest je omjer vjerojatnosti da imamo tu bolest i da ju nemamo i koristimo

formulu (1.13). Na primjer, ako je vjerojatnost da imamo bolest 0.6, onda šansa da dobijemo bolest iznosi $O = \frac{0.6}{1-0.6} = 1.5$.
Zašto su nam potrebne šanse? Zbog razumnijeg višestrukog uspoređivanja. Na primjer, ako je vjerojatnost da osoba X ima bolest 0.3, a osoba Y 0.6, zaključujemo da je vjerojatnost osobe Y duplo veća od vjerojatnosti osobe X. S druge strane, ako je vjerojatnost da osoba X ima bolest 0.9, onda bi vjerojatnost da osoba Y ima bolest trebala iznositi 1.8 što je nemoguće pošto najveća vrijednost vjerojatnosti je 1. Zato je za uspoređivanje razumnije koristiti šanse. Primjerice, ako imamo $p = 0.7$ to povlači da šansa iznosi $O = \frac{0.7}{0.3} \approx 2.33$. Dvostruka vrijednost te šanse je 4.66, pa vraćanjem na vjerojatnost dobivamo $p \approx 0.82$. Time dolazimo do omjera šansi, široko korištenoj mjeri koja mjeri odnos dviju dihotomnih varijabli. Omjer šansi, u ovom primjeru, je šansa rizičnih kroz šansa nerizičnih. Uvedimo dvije dodatne oznake, neka je $p_{\text{rizični}} = \frac{a}{a+b}$ vjerojatnost da osoba koja je bolesna pije alkohol i $p_{\text{nerizični}} = \frac{c}{c+d}$ vjerojatnost da osoba koja je bolesna ne pije alkohol. Znači, šansa rizičnih je omjer vjerojatnosti onih koji su izloženi rizičnom faktoru imaju bolest i koji nemaju bolest, analogno vrijedi i za šansu nerizičnih. Odnosno, $O_{\text{rizični}} = \frac{p_{\text{rizični}}}{1-p_{\text{rizični}}}$ i $O_{\text{nerizični}} = \frac{p_{\text{nerizični}}}{1-p_{\text{nerizični}}}$. Dakle dobivamo da je omjer šanse (OR)

$$OR = \frac{O_{\text{rizični}}}{O_{\text{nerizični}}} = \frac{\frac{p_{\text{rizični}}}{1-p_{\text{rizični}}}}{\frac{p_{\text{nerizični}}}{1-p_{\text{nerizični}}}} \implies OR = \frac{\frac{a}{b}}{\frac{c}{d}} = \frac{ad}{bc}.$$

Tablica 1.3: Tablica kontingencije pijenja alkohola i stanja bolesti

		Stanje		**Ukupno**
		Bolestan	Zdrav	
Rizičan faktor	Pije alkohol	21	10	31
	Ne pije alkohol	14	25	39
Ukupno		35	35	70

Primjer 1.4.1. *Prema tablici 1.3 imamo:*
$O_{pije\ alkohol} = \frac{21}{10} = 2.1 \implies$ *osobe koji piju alkohol imaju 2.1 puta veću šansu da dobiju bolest nego da ju ne dobiju.*
$O_{ne\ pije\ alkohol} = \frac{14}{25} = 0.56 \implies$ *osobe koje ne piju alkohol imaju 0.56 puta veću šansu da dobiju bolest nego da ju ne dobiju.*
Zaključak: osobe koje ne piju alkohol imaju manju šansu dobivanja bolesti naspram osoba koje piju alkohol.
$OR = \frac{21 \cdot 25}{10 \cdot 14} = 3.75 \implies$ *osobe koje piju alkohol imaju 3.75 puta veći omjer šanse da dobiju bolest.*

Vidjet ćemo da su omjeri šansi direktno povezani s parametrima u logit modelu. [2, 8]

1.5 Logit model

Logit model se još zove i model logističke regresije. Kao što smo rekli i prije, glavni problem s linearnim vjerojatnosnim modelom je taj što su vjerojatnosti ograničene s 0 i 1, a dok je linearna funkcija neograničena. Rješenje je transformirati vjerojatnost tako da više ne bude ograničena.

Tranformiranjem vjerojatnosti u šanse pomoću prigodne jednadžbe (1.13) mičemo gornju granicu. Ako djelujemo s logaritmom na šansu, time ćemo maknuti donju granicu. Izjednačimo dobiveni rezultat s linearnom funkcijom varijabli poticaja,tako dobivamo logit model. Za k varijabli poticaja i $i = 1, 2, \ldots, n$ ponavljanja, model glasi

$$\log\left[\frac{p_i}{1-p_i}\right] = \text{logit}\,(p_i) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} \tag{1.14}$$

gdje je p_i vjerojatnost da $y_i = 1$. Izraz na lijevoj stranici izraza (1.14) zovemo *logit* ili *log-šansa*. (Napomenimo da se pri korištenju logaritma smatramo korištenje prirodnog logaritma).

Za razliku od linearnog modela, u logit modelu se ne pojavljuje pojam slučajne pogreške. To ne znači da je model deterministički zato jer i dalje ima prostora za slučajnu varijaciju u vjerojatnosnom odnosu između p_i i y_i. Eksponencijalna funkcija je neprekidna funkcija, te primijenimo li ju na (1.14) dobivamo sljedeće

$$p_i = \frac{\exp(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik})}{1 + \exp(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik})} \tag{1.15}$$

pri čemu smo koristili svojstvo $\log(e^x) = x$. Podijelimo li brojnik i nazivnik u razlomku s njegovim brojnikom dobivamo

$$p_i = \frac{1}{1 + \exp(-\beta_0 - \beta_1 x_{i1} - \cdots - \beta_k x_{ik})}. \tag{1.16}$$

U dobivenoj jednadžbi p_i poprima vrijednost 0 ili 1 za bilo koju vrijednost $\beta_0, \beta, \ldots, \beta_k$ i x.

Ako imamo jednu varijablu x sa $\beta_0 = 0$ i $\beta_1 = 1$, tada jednažba (1.16) izgleda $p = \frac{1}{1+e^{-x}}$ te njezin graf se zove S-krivulja (eng. *S-shaped curve*). Kako x postaje veći ili manji, tako se p približava 1 ili 0 ali nikad nije jednaka tim granicama. Primjer S-krivulje možemo vidjeti u poglavlju 3. [2]

1.6 Procjena parametara u modelu logističke regresije

Za razliku dok kod metode najmanjih kvadrata minimiziramo kvadrirane reziduale kod logističke regresije koristimo metodu maksimalne vjerodostojnosti (eng. *Maximum likelihood*, u daljnjem tekstu ML). Kod ML tražimo najmanje moguće odstupanje (eng. *Deviance*) između opaženih i predikivnih vrijednosti. Koristimo iterativne računalne metode sve dok ne dobijemo najmanje moguće odstupanje, te dobiveno rješenje zovemo *Likelihood ratio* ili *-2LogLikelihood* ili DEVIANCE.

Pretpostavimo da imamo n nezavisnih observacija $(\mathbf{x}_i, y_i)$ za $i = 1, 2, \ldots, n$ tako da je y_i slučajna varijabla koja poprima vrijednosti 0 ili 1 i $\mathbf{x}_i = [1\ x_{i1}\ \ldots\ x_{ik}]'$ je vektor varijabli poticaja (1 je za slobodni član). Vektorski prikaz je od pomoći radi ljepše preglednosti jednadžbi. Ukoliko je p_i vjerojatnost da $y_i = 1$, onda pretpostavljamo da su podaci generirani s logit model koji kaže

$$p(x) = \frac{1}{1 + e^{-\beta \mathbf{x}_i}}, \tag{1.17}$$

što je ekvivalentno izrazu (1.16) i gdje je $\beta = (\beta_0, \beta_1, \ldots, \beta_k)$. Slučajna varijabla y_i ima Bernoulijevu razidobu s parametrom p_i za $i = 1, 2, \ldots, n$, pa imamo

$$\mathbb{P}(y_i = k) = \begin{cases} p_i & , k = 1 \\ 1 - p_i & , k = 0 \end{cases} \implies f(y_i|p_i) = p_i^{y_i}(1 - p_i)^{1-y_i}. \tag{1.18}$$

Definicija 1.6.1. *Neka je $(x_1, \ldots, x_n)$ opaženi uzorak za slučajnu varijablu X s gustoćom $f(x|\theta)$, gdje je $\theta = (\theta_1, \ldots, \theta_k) \in \Theta \subseteq \mathbb{R}^k$ nepoznati parametar.*
Definiramo ***funkciju vjerodostojnosti*** *$L : \Theta \to \mathbb{R}$ sa*

$$L(\theta) := \prod_{i=1}^{n} f(x_i|\theta), \theta \in \Theta. \tag{1.19}$$

Vrijednost $\hat{\theta} = \hat{\theta}(x_1, \ldots, x_n) \in \Theta$ za koju je

$$L(\hat{\theta}) = \max_{\theta \in \Theta} L(\theta) \tag{1.20}$$

zovemo ***procjena metodom maksimalne vjerodostojnosti****. Statistika $\hat{\theta} = \hat{\theta}(X_1, \ldots, X_n)$ je* ***procjenitelj metodom maksimalne vjerdostojnosti*** *ili kraće* ***MLE****.*

Koristeći definiciju 1.6.1 i nezavisnost od $y = (y_1, \ldots, y_n)$ dobivamo

$$L(\beta) = \prod_{i=1}^{n} p_i^{y_i}(1 - p_i)^{1-y_i} = \prod_{i=1}^{n} \left(\frac{p_i}{1 - p_i}\right)^{y_i} (1 - p_i). \tag{1.21}$$

Princip maksimalne vjerodostojnosti je da procjena β maksimizira L(β), da bi nam lakše išao račun logaritmiramo izraz (1.21).

$$\ln L(\beta) = \sum_{i=1}^{n} y_i \ln\left(\frac{p_i}{1-p_i}\right) + \sum_{i=1}^{n} \ln(1-p_i) \tag{1.22}$$

Općenito je lakše raditi s logaritmom funkcije vjerodostojnosti zbog toga jer logaritam umnoška je suma logaritama i eksponenti postaju koeficijenti. Pošto je logaritam rastuća funkcija, što god maksimizira logaritam također će maksimizirati i originalnu funkciju. Uvršavajući naš izraz za logit model (1.17) u jednadžbu (1.22) dobivamo

$$\ln L(\beta) = \sum_{i=1}^{n} \beta \mathbf{x}_i y_i - \sum_{i=1}^{n} \ln(1 + e^{\beta \mathbf{x}_i}) \tag{1.23}$$

Idući korak je izabrati najveći mogući β koji maksimizira jednadžbu (1.23). Postoje različite metode za maksimiziranje ovakvih funkcija, jedan dobro poznati pristup je derivacija funkcije po β te izjednačiti s 0 i pronaći rješenje za β. Derivacijom jednadžbe (1.23) i izjednačavanje s 0 dobivamo idući izraz

$$\begin{aligned} \frac{\partial \ln L(\beta)}{\partial \beta} &= \sum_{i=1}^{n} \mathbf{x}_i y_i - \sum_{i=1}^{n} \frac{\mathbf{x}_i e^{\beta \mathbf{x}_i}}{1 + e^{\beta \mathbf{x}_i}} \\ &= \sum_{i=1}^{n} \mathbf{x}_i y_i - \sum_{i=1}^{n} \mathbf{x}_i \frac{1}{1 + e^{-\beta \mathbf{x}_i}} \\ &= \sum_{i=1}^{n} \mathbf{x}_i y_i - \sum_{i=1}^{n} \mathbf{x}_i \hat{y}_i = 0 \end{aligned} \tag{1.24}$$

gdje je $\hat{y}_i = \frac{1}{1+e^{-\beta \mathbf{x}_i}}$ prediktivna vjerojatnost od y_i za danu vrijednost $\mathbf{x}_i$. Zbog toga jer je $\mathbf{x}_i$ vektor, jednadžba (1.24) zapravo ima $k + 1$ jednadžbi, za svaki element od β. Jednadžbe (1.24) su nelinearne po β pa ih rješavamo jednom od iterativnih metoda. Najčešća iterativna metoda je Newton-Raphson algoritam, koju ćemo sada ukratko objasniti. Neka je $U(\beta)$ vektor prve dervacije od $\ln L$ po β i neka je $I(\beta)$ matrica druge derivacije od $\ln L$ po β. Imamo

$$\begin{aligned} U(\beta) &= \frac{\partial \ln L(\beta)}{\partial \beta} = \sum_{i=1}^{n} \mathbf{x}_i y_i - \sum_{i=1}^{n} \mathbf{x}_i \hat{y}_i \\ I(\beta) &= \frac{\partial^2 \ln L(\beta)}{\partial \beta^2} = -\sum_{i=1}^{n} \mathbf{x}_i^2 \hat{y}_i (1 - \hat{y}_i) \end{aligned}$$

gdje $U(\beta)$ još zovemo gradijent, a $I(\beta)$ Hesseova matrica. Newton-Raphson algoritam tada je

$$\beta^{j+1} = \beta_j - I^{-1}(\beta^j) U(\beta^j) \tag{1.25}$$

gdje je I^{-1} inverz od I. U praksi, trebamo skup početnih podataka za β^0. U našem slučaju, u poglavlju 3 koristimo programski jezik SAS, procedura LOGISTIC početne vrijednosti za β^0 postavlja sve koeficijente na 0. Početne vrijednosti ubacujemo u desnu stranu jednadžbe (1.25) i time dobivamo rezultat za prvu iteraciju β^1. Vrijednosti za β^1 ponovo ubacujemo u desnu stranu jednadžbe (1.25), te nakon što su izračunate prva i druga derivacija, dobivamo rezultat za drugu iteraciju β^2. Ovaj postupak ponavljamo dok makismalna promjena u svakom procijenjenom parametru iz jednog koraka u drugi je manja od nekog zadanog kriterija. Ako je apsolutna vrijednost trenutnog procijenjenog parametra β^j manja ili jednaka od 0.01, onda je zadani kriterij za konvergenciju

$$\left|\beta^{j+1} - \beta^j\right| < 0.0001.$$

Ako je apsolutna vrijednost trenutnog procijenjenog parametra β^j veća od 0.01, onda je zadani kriterij

$$\left|\frac{\beta^{j+1} - \beta^j}{\beta^j}\right| < 0.0001.$$

Nakon što je pronađeno rješenje za $\hat{\beta}$, dobiveni produkt u Newton-Raphsonovom algoritmu je procjena kovarijacijske matrice koeficijenta, što je jednostavno $-I^{-1}(\hat{\beta})$. Procjena standardnih pogrešaka koeficijenta dobivene su uzimanje drugog korijena elemenata koji se nalaze na glavnoj dijagonali ove matrice. [2, 8]

Problem konvergencije

Kao što smo spomenuli u prethodnom poglavlju, ML za procjenu logističkog modela je iterativna metoda koja podrazumijeva uzastopne aproksimacije. Kada je promjena koeficijenata u dvije uzastopne iteracije manja od nekog zadanog kriterija, računanje prestaje i algoritam daje informaciju da je konvergencija zadovoljena. U većini slučajeva, uvjet konvergencije je zadovoljen. U nekim slučajevima iterativna metoda se zaustavlja, ali konvergencija nije zadovoljena. U slučajevima kada konvergencija nije zadovoljena, problem nije u postavljenom limitu broja iteracija, već procjena ML ne postoji. Ako postoji neka linearna kombinacija nezavisnih varijabli koja savršeno predviđa zadanu varijablu odaziva, onda neće biti zadovoljen uvjet konvergencije te je to jedan od češćih razloga ne zadovoljavanja uvjeta konvergencije.

1.7 Testiranje adekvatnosti modela (eng. *Goodness of fit*)

Neka su $\hat{y}$ ML procjena za y i $\hat{p}$ ML procjena za p. Nas zanima koliko se razlikuju $\hat{y}$ i y, pošto želimo da su procjenjene vrijednosti što bliže opažanim vrijednostima. Za testiranje adekvatnosti modela u logističkoj regresiji koristi se devijacija (eng. *Deviance*, nadalje u tekstu u oznaci D).

U izračunu devijacije za logit model, maksimalni model se često naziva *zasićenim* modelom. Zasićeni model ima jedan parametar za svaku predviđenu vjerojatnost i time se dobiva savršeno prilagođavanje podacima. Usporedimo li model s k parametara, $1 < k < n$, sa zasićenim modelom, možemo vidjeti koliko je taj model u stvari dobar. Devijacija ima istu ulogu kao i suma kvadrata reziduala u linearnom modelu. Devijaciju računamo na sljedeći način:

$$\begin{aligned} D &= -2\ln\left[\frac{\text{ML modela sa } k \text{ parametara}}{\text{ML zasićenog mode}}\right] \\ &= -2\sum_{i=1}^{n}\left[y_i \ln\left(\frac{\hat{p}(x_i)}{y_i}\right) + (1-y_i)\ln\left(\frac{1-\hat{p}(x_i)}{1-y_i}\right)\right] \approx \chi^2. \end{aligned} \tag{1.26}$$

Omjer ML modela sa k parametara i ML zasićenog modela zovemo *omjer vjerodostojnosti*.

Za potrebe procjene značajnosti nezavisne varijable u modelu, uspoređujem vrijednosti D s i bez nezavisnih varijabli u jednadžbi, te se ta statistika zove G statistika.

$G = D$(model bez nezavisnih varijabli, samo slobodan član)$-D$(model s k nezavisnih varijabli)

Kako su obje vrijednosti D jednake ML zasićenog modela, G možemo izraziti kao:

$$G = -2\ln\left[\frac{\text{ML modela bez nezavisnih varijabli}}{\text{ML modela s } k \text{ nezavisnih varijabli}}\right] = -2\ln\left[\frac{L(\beta_0)}{L(\beta_0,\beta_1,\ldots,\beta_k)}\right] \approx \chi^2(k).$$

[3]

1.8 *ROC* krivulja

ROC (eng. *Receiver Operating Characteristic*) krivulja je grafički prikaz odnosa mjera valjanosti dijagnostičkog testa.. Valjanost dijagnostičkog testa je složeni pokazatelj i ima dvije komponente: osjetljivost i specifičnost. Osjetljivost testa je proporcija dobro detektiranih ženskih glasova od sveukupnog broja ženskih glasova, a

specifičnost testa je proporcija muških glasova koji su dobro detektirani kao muški glasovi, od ukupnog broja muških glasova. Analiza osjetljivosti i specifičnosti testa ovisno o postavljanju granice koja odvaja "test-pozitivne" od "test-negativnih", naziva se *ROC* analiza. Upravo jedan od efikasnijih načina da se prikaže veza između osjetljivosti i specifičnosti testa je takozvana *ROC* krivulja. Glavna ideja takve krivulje je prikaz odnosa proporcija lažno pozitivnih (1-specifičnost) i stvarno pozitivnih (osjetljivost).

Površina ispod *ROC* krivulje predstavlja prediktivnu snagu modela (pod oznakom c, eng. *Concordance Index*), te se još zove c-statistika. Vrijednosti c-statistike variraju između 0 i 1.
Formula za c-statistiku glasi:

$$c = \frac{nc + 0.5(t - nc - nd)}{t} \tag{1.27}$$

gdje t predstavlja broj parova s različitim vrijednostima odgovora, nc je broj usklađenih (eng. *concordant*) parova i nd broj neusklađenih (eng. *disconcordant*) parova. Odnosno, definirajmo da se događaj dogodio s 1, a da se nije dogodio s 0. Za par observacija s različitim odgovorima, kažemo da su usklađene ako observacija koja ima više rangirani odgovor (npr. 2 "događaj se ne dogodi"), ima nižu prediktivnu vjerojatnost da se događaj dogodi od opservacije s niže rangiranim odgovorom (npr. 1 "događaj se dogodi"). Za par observacija s različitim odgovorima, kažemo da su neusklađene ako observacija koja ima više rangirani odgovor, ima višu prediktivnu vjerojatnost da se događaj dogodi od observacije s niže rangiranim odgovorom. Ako par observacija nije ni usklađen ni neusklađen, kažemo da je odgovor jednak (eng. *tie*). [8]

Poglavlje 2

Klasifikacijska i regresijska stabla odlučivanja

2.1 Općenito

Klasifikacijska i regresijska stabla odlučivanja (eng. *Classification and Regression Tree*, u daljnjem tekstu CART analiza) je jednostavan ali i moćan analitički alat koji pomaže odrediti "značajne" varijable u određenom skupu podataka, što može pomoći u konstrukciji dobrog modela. CART analiza je postala sve popularnija, a posebno je vrijedna u multidisciplinarnim područjima. To je neparametarska statistička metoda koja se razlikuje ovisno o definiciji izlazne varijable. Pozadina CART analize je matematički identična i dobro poznatim regresijskim tehnikama, ali predstavlja podatke na način koji se lako tumači onima koji nisu dobro upućeni u statistički analizu. Na taj način, CART analiza predstavlja sofisticiranu vizualnu sliku odnosa varijabli u skupu podataka i može se koristiti kao prvi korak pri izgradnji informativnog modela ili za završnu vizualizaciju bitnih asocijacija. [5]

Postoje dvije vrste stabla odlučivanja:

- **Klasifikacijsko stablo** - gdje je zavisna varijabla kategorijska (najčešće binarna) te se stablo koristi za prepoznavanje "klase" unutar koje bi se vjerojatno pojavila zavisna varijabla
- **Regresijsko stablo** - gdje je zavisna varijabla neprekidna te se stablo koristi za predviđanje njezine vrijednosti

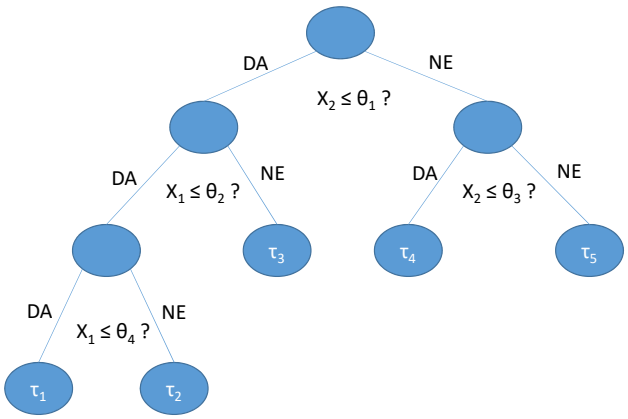
2.2 Klasifikacijsko stablo

Izgradnja stabla

Algoritam za izgradnju stabla koristi podatke kako bi razdvojio prediktorne skupove (odnosno, skup svih mogućih kombinacija vrijednosti prediktorne varijable) u područja koja se ne preklapaju. Ova područja odgovaraju terminalnim (ili konačnim) čvorovima stabla koja se inače zovu *listovi* stabla. Svako područje je opisano skupom pravila, a ta pravila se koriste kako bi se nova observacija dodijelila određenom području. Za klasifikacijsko stablo, prediktorna vrijednost za ovu observaciju je najčešća razina varijable odaziva u tom području.

Dijeljenje stabla se vrši rekurzivno, počevši od svih observacija, koje predstavljaju čvor na vrhu stabla (odnosno *korijen* stabla). Algoritam razdvaja roditeljski čvor u dva ili više dječjih čvorova tako da odgovori unutar svakog dječjeg područja budu što sličniji. Proces dijeljenja se zatim ponavlja za svaki dječji čvor, te se rekurzija nastavlja dok kriterij zaustavljanja (eng. *stopping criterion*) ne bude zadovoljen i stablo je potpuno izgrađeno.

U svakom koraku, dijeljenje se određuje pronalaženjem najbolje prediktorne varijable i najboljeg mjesta za dijeljenje (eng. *cutpoint*) (ili takvog skupa) pa se na taj način observacije u roditeljskim čvorovima dodjeljuju dječjim čvorovima. U sljedećim poglavljima ćemo bolje pojasniti kriterij dijeljenja i strategiju dijeljenja.



Slika 2.1: Primjer klasifikacijskog stabla gdje su X_1 i X_2 ulazne varijable i $\theta_i, i = 1, 2, 3, 4$ vrijednosti

Primjer 2.2.1. *Razmotrit ćemo jednostavan primjer rekurzivnog particioniranja koje uključuje dvije ulazne varijable, X_1 i X_2. Pretpostavimo da stablo izgleda kao što je prikazano na slici 2.1, također bez smanjenja općenitosti možemo pretpos-*

taviti $\theta_1 < \theta_3$ i $\theta_2 < \theta_4$. Moguće faze ovog stabla su: (1) Vrijedi li $X_2 \leq \theta_1$? Ako je odgovor DA, onda pratimo lijevu granu tog čvora, a ako NE onda pratimo desnu granu. (2) Ako je odgovor na (1) pozitivan, onda se pitamo iduće pitanje, vrijedi li $X_1 \leq \theta_2$? Pozitivan odgovor nas vodi do novog pitanja, vrijedi li $X_1 \leq \theta_4$? Odgovor DA doveo nas je do lista τ_1 sa odgovarajućim područjem $R_1 = \{X_1 \leq \theta_2, X_2 \leq \theta_1\}$, a odgovor NE nas vodi do lista τ_2 sa pripadnim područjem $R_2 = \{\theta_4 < X_1 \leq \theta_2, X_2 \leq \theta_1\}$. Imamo li negativan odgovor na (2), dolazimo do lista τ_3 sa pripadnim područjem $R_3 = \{X_1 > \theta_2, X_2 \leq \theta_1\}$. Vraćamo se na početak, ako smo na (1) odgovorili negativno, postavljamo novo pitanje, vrijedi li $X_2 \leq \theta_3$? Pozitivan odgovor nas vodi do lista τ_4 sa pripadnim područjem $R_4 = \{\theta_1 < X_2 \leq \theta_3\}$, a negativan odgovor do lista τ_5 sa pripadnim područjem $R_5 = \{X_2 > \theta_3\}$.

Kriterij dijeljenja

U svakom čvoru, algoritam treba odlučiti koja varijabla je najpogodnije za dijeljenje. Trebaju se razmotriti sve mogućnosti dijeljenja za svaku varijablu prisutnu u tom čvoru, prebrojati sva moguća dijeljenja te ih procijeniti i onda odrediti koje dijeljenje je najbolje u nekom određenom smislu. Prije opisa pravila dijeljenja moramo napraviti razliku između ordinalnih i nominalnih varijabli.

Za neprekidnu varijablu, broj mogućih dijeljenja u određenom čvoru je jedan manje od broja njenih različitih observacija. Pretpostavimo da za neku kategorijsku varijablu imamo P kategorija, odnosno neka su te kategorije $l_1, l_2, \ldots, l_P$. Sa $\mathcal{S}$ označim skup svih mogućih dijeljenja za čvor čija varijabla je skup svih podskupova od $\{l_1, l_2, \ldots, l_P\}$. Općenito imamo $2^{P-1} - 1$ različitih dijeljenja za skup $\mathcal{S}$.

Da bismo mogli izabrati najbolje dijeljenje među svim varijablama, prvo moramo odabrati najbolje dijeljenje za određenu varijablu. Neka je X matrica promatranih podataka koja se sastoji od N podataka i M varijabli i neka je Y N-dimenzionalni zavisni vektor s najviše K klasa. Sa τ_R, τ_L i τ_D označio redom čvorove roditelja, te lijevo i desno dijete.

Neka je x_j varijabla $j = 1, 2, \ldots, M$, a x_j^R najbolja vrijednost varijable x_j koja dijeli čvor. Roditeljska nečistoća je konstantna za svaku varijablu za koju vrijedi $x_j \leq x_j^R$, dok je maksimalna homogenost djece ekvivaletna maksimalnoj vrijednosti funkcije nečistoće, imamo

$$\Delta i(\tau) = i(\tau_R) - \mathbb{E}[i(\tau_K)] \tag{2.1}$$

gdje je τ_K lijevo i desno dijete roditeljskog čvora τ_R.

Pod pretpostavkom da su p_L i p_D vjerojatnosti lijevog i desnog čvora, adekvatnost modela (eng. *Goodness of fit*) u čvoru τ je dana dijeljenjem čvora na dva dijeteta formulom

$$\Delta i(\tau) = i(\tau_R) - p_L i(\tau_L) - p_D i(\tau_D). \tag{2.2}$$

Problem maksimalne homogenosti funkcije se rješava kao

$$\arg \max_{x_j \leq x_j^R, j=1,2,\ldots,M} [i(\tau_R) - p_L i(\tau_L) - p_D i(\tau_D)]. \tag{2.3}$$

Nadalje, neka su $\Pi_1, \ldots, \Pi_K$, $K \geq 2$ klase. Za čvor τ definiramo funkciju nečistoće čvora (eng. *node impurity function*) $i(\tau)$ sa

$$i(\tau) = \phi(p(1|\tau), \ldots, p(K|\tau)), \tag{2.4}$$

gdje je $p(k|\tau)$ procjenitelj za $\mathbb{P}(\mathbf{X} \in \Pi_k|\tau)$, što je uvjetna vjerojatnost da observacija $\mathbf{X}$ unutar klase Π_k upada u čvor τ. Funkcija ϕ je simetrična funkcija definirana na skupu svih K-torki vjerojatnosti $(p_1, \ldots, p_K)$ sa jediničnom sumom.
Neki poznatiji primjeri takvih funkcija ϕ su:

- *Funkcija entropije*

$$i(\tau) = -\sum_{k=1}^{K} p(k|\tau) \ln p(k|\tau) \tag{2.5}$$

- *Gini indeks ili Gini pravilo*

$$i(\tau) = \sum_{k=1}^{K} p(k|\tau)(1 - p(k|\tau)) = \sum_{k \neq k'} p(k|\tau)p(k'|\tau) = 1 - \sum_{k=1}^{K} (p(k|\tau))^2. \tag{2.6}$$

U slučaju kada imamo dvije klase, gornji izrazi se reduciraju na

- $i(\tau) = -p \ln p - (1 - p) \ln(1 - p)$
- $i(\tau) = 2p(1 - p)$

gdje je $p = p(1|\tau)$. [4]

Odabir najboljeg dijeljenja za varijablu

Pretpostavimo, u čvoru τ, primijenimo dijeljenje s tako da udio p_L observacija ode u čvor lijevog dijeteta τ_L, a preostali udjel p_D ode u čvor desnog dijeteta τ_D. Na primjer, pretpostavimo da imamo skup podataka i varijablu odaziva Y koja poprima vrijednosti 0 ili 1. Pretpostavimo da jedno od mogućih dijeljenja varijable poticaja X_j je $X_j \leq c$ ili $X_j > c$, gdje je c neka vrijednost od X_j. To možemo ljepše zapisati u 2×2 tablici, koja je prikazana u tablici 2.1

Razmotrimo prvo, roditelji čvor τ_R. Za mjeru nečistoće koristimo funkciju entropije 2.5. Uvedimo nove oznake n_{ab}, n_{a+}, n_{+a} i n_{++} tako da $a, b \in \{1, 2\}$ za koje vrijedi

- $n_{a+} = n_{a1} + n_{a2}$
- $n_{+a} = n_{1a} + n_{2a}$
- $n_{++} = n_{a+} + n_{a+} = n_{+a} + n_{+a}$.

Ostale veličine procjenjujemo iz tablice 2.1 na način da gledamo broj podataka iz skupa koji zadovoljavaju uvjete u križanju određenog retka i stupca za promatranu varijablu. Procijenimo p_L sa n_{+1}/n_{++} i p_D sa n_{+2}/n_{++} i funkcija nečistoće za roditeljski čvor

$$i(\tau_R) = -\left(\frac{n_{+1}}{n_{++}}\right)\ln\left(\frac{n_{+1}}{n_{++}}\right) - \left(\frac{n_{+2}}{n_{++}}\right)\ln\left(\frac{n_{+2}}{n_{++}}\right). \tag{2.7}$$

Napomena 2.2.2. *$i(\tau)$ je potpuno nezavisna o vrsti predloženog dijeljenja.*

Analogno za dječje čvorove τ_L i τ_D imamo

- za $x_j \leq x_j^R$ procijenimo p_L sa n_{11}/n_{1+} i p_D sa n_{12}/n_{1+}
- za $x_j > x_j^R$ procijenimo p_L sa n_{21}/n_{2+} i p_D sa n_{22}/n_{2+}.

Tada dobijemo:

$$\begin{aligned} i(\tau_L) &= -\left(\frac{n_{11}}{n_{1+}}\right)\ln\left(\frac{n_{11}}{n_{1+}}\right) - \left(\frac{n_{12}}{n_{1+}}\right)\ln\left(\frac{n_{12}}{n_{1+}}\right). \\ i(\tau_D) &= -\left(\frac{n_{21}}{n_{2+}}\right)\ln\left(\frac{n_{21}}{n_{2+}}\right) - \left(\frac{n_{22}}{n_{2+}}\right)\ln\left(\frac{n_{22}}{n_{2+}}\right). \end{aligned} \tag{2.8}$$

Tablica 2.1: 2×2 tablica za dijeljenje varijable x_j, gdje varijabla odaziva poprima vrijednosti 0 i 1

	1	**0**	**ZBROJ**
$\mathbf{x_j \leq x_j^R}$	n_{11}	n_{12}	n_{1+}
$\mathbf{x_j > x_j^R}$	n_{21}	n_{22}	n_{2+}
ZBROJ	n_{+1}	n_{+2}	n_{++}

Na kraju računamo vrijednost funkcije (2.2) za promatranu varijablu, što provjerimo za svaku vrijednost iz skupa te varijable formulom (2.3) pa odaberemo maksimalnu vrijednost funkcije $i(\tau)$. Navedeni proces traženja maksimuma funkcije radimo za svaki skup varijabli posebno u skupu kojeg promatramo. Zatim iz novo dobivenog skupa maksimalnih vrijednosti tražimo maksimalnu vrijednost koja će odrediti varijablu x_j^R za $j = 1, 2, \ldots, M$, odnosno varijablu s najboljim uvjetom za dijeljenje čvora. [4]

Procjena greške grupiranja

Procjena greške grupiranja $r(\tau)$ za observacije u čvoru τ je dana formulom

$$r(\tau) = 1 - \max_k p(k|\tau), \tag{2.9}$$

dana formula se u slučaju dvije klase (oznaka klase je k) reducira na $r(\tau) = 1 - \max(p, 1-p) = min(p, 1-p)$.

Neka je T oznaka stabla i neka je $\tilde{T} = \{\tau_1, \tau_2, \ldots, \tau_L\}$ skup svih konačnih čvorova od T. Tada možemo procijeniti pravu pogrešku grupiranja za T sa

$$R(T) = \sum_{\tau \in \tilde{T}} R(\tau)\mathbb{P}(\tau) = \sum_{l=1}^{L} R(\tau_l)\mathbb{P}(\tau_l) \tag{2.10}$$

gdje je $\mathbb{P}(\tau)$ vjerojatnost da observacija upadne u čvor τ. Procijenimo li $\mathbb{P}(\tau_l)$ s udjelom $p(\tau_l)$ svih observacija koje upadnu u čvor τ_l, tada je izmjenjena procjena od $R(T)$ dana s

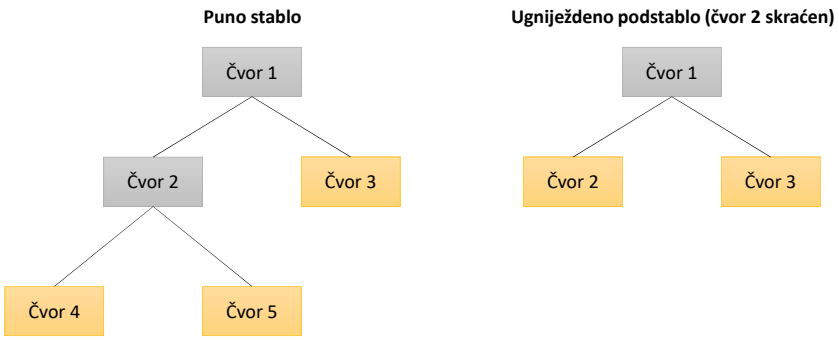
$$R^{re}(T) = \sum_{l=1}^{L} r(\tau_l)p(\tau_l) = \sum_{l=1}^{L} R^{re}(\tau_l) \tag{2.11}$$

gdje je $R^{re}(\tau_l) = r(\tau_l)p(\tau_l)$ za $l = 1, \ldots, L$. Drugim riječima, procjenom greške grupiranja zapravo vidimo koliko nam se razlikuju procijenjene vrijednosti od stvarnih. [4]

Skraćivanje stabla

Statističar L.Breiman koji je pridonio razvoju klasifikacijskih i regresijskih stabala odlučivanja ima filozofiju da se izgadnja konačnog stabla sastoji od toga da stablo raste do "maksimalne veličine" te onda skratimo grane dok ne dobijemo "pravu veličinu" stabla. Skraćeno stablo je podstablo (eng. *subtree*) originalnog velikog stabla. Često je "skupo" procijeniti pogrešku na svim mogućim podstablima punog stabla. Praktičnija strategija je usredotočiti se na niz ugniježdenih stabala dobivenih uzastopnim skraćivanjem listova stabla.

Slika 2.2: Primjer punog stabla i skraćenog podstabla



Primjer 2.2.3. *Slika 2.2 prikazuje primjer skraćivanja u kojem su listovi od Čvora 2 (Čvor 4 i Čvor 5) maknuti kako bismo dobili ugniježđeno podstablo punog stabla. U ugniježđenom podstablu, Čvor 2 je sada list i sadrži sve observacije koje su prije sadržavali Čvor 4 i Čvor 5. Ovaj proces se ponavlja dok ne dođemo samo do korijena stabla.*

Postoji puno različitih metoda za skraćivanje stabla. Te metode odnose se i na način odabira čvorova koji će se izrezati kako bi stvorio niz podstabala i kako zatim odabrati optimalno podstablo iz tog niza kako bismo dobili završno stablo. Jedna od poznatijih metoda za skraćivanje stabala je posljedična složenost skraćivanja stabla (eng. *cost-complexity prunning tree*).

Algoritam skraćivanja se sastoji od:

1. Napravimo veliko stablo, $T_{\max}$, s maksimalnim skupom čvorova tako da dijelimo čvorove sve dok svaki sadrži manje od $n_{\min}$ observacija;
2. Izračunamo procjenu od $R(\tau)$ u svakom čvoru $\tau \in T_{\max}$;
3. Skraćujemo stablo $T_{\max}$ prema korijenu stabla tako da u svakom koraku skraćivanja procjena od $R(T)$ bude minimalna.

Neka je $\alpha \geq 0$ parametar složenosti (eng. *complexity parameter*). Za bilo koji čvor $\tau \in T$ stavimo

$$R_\alpha(\tau) = R^{re}(\tau) + \alpha. \tag{2.12}$$

Iz gornjeg izraza (2.12) definiramo mjeru posljedične složenosti skraćivanja stabla (eng. *cost-complexity prunning measure*) kao

$$R_\alpha(T) = \sum_{l=1}^{L} R_\alpha(\tau_l) = R^{re}(T) + \alpha|\tilde{T}|, \tag{2.13}$$

gdje je $|\tilde{T}| = L$ broj konačnih čvorova u podstablu T od $T_{\max}$. Na $\alpha|\tilde{T}|$ gledamo kao na neku vrstu "kaznenog" izraza zbog veličine stabla, odnosno $R_\alpha(T)$ može kazniti $R^{re}(T)$ za generiranje prevelikog stabla. Za svaki α odabremo ono podstablo $T(\alpha)$ od $T_{\max}$ koje minimizira $R_\alpha(T)$

$$R_\alpha\left(T(\alpha)\right) = \min_T R_\alpha(T). \tag{2.14}$$

Ako $T(\alpha)$ zadovoljava (2.14), onda se zove minimizirano podstablo (tj. optimalno skraćeno podstablo) od $T_{\max}$. Za bilo koji α možemo pronaći više od jednog minimizirajućeg podstabla od $T_{\max}$. Vrijednost α određuje veličinu stabla. Znači, kada je α jako mali, kaznena vrijednost će također biti mala, dok će veličina minimizirajućeg podstabla koje je određeno s $R^{re}\left(T(\alpha)\right)$ biti velika.

Breiman je pokazao kako za svaku vrijednost α, postoji podstablo od T koje minimizira posljedičnu složenost. Kada je $\alpha = 0$ tada je to puno stablo T_0. Kako α raste, odgovarajuće podstablo postaje manje i podstabla su zapravo ugniježdena. Onda, za neku vrijednost α korijen stabla ima minimalnu posljedičnu složenost za svaki α koji je veći ili jednak od te vrijednosti. Zbog toga jer postoji konačan broj podstabala, svako podstablo odgovara nekom intervalu vrijednosti od α, tj

$$\begin{aligned}
[0, \alpha_1) &= \text{ interval gdje } T_0 \text{ (puno stablo) ima najmanju posljedičnu složenost} \\
[\alpha_1, \alpha_2) &= \text{ interval gdje } T_1 \text{ ima najmanju posljedičnu složenost} \\
&\vdots \\
[\alpha_m, \infty) &= \text{ interval gdje } T_m \text{ (korijen) ima najmanju posljedičnu složenost}
\end{aligned} \tag{2.15}$$

Kako doći od $T_{\max}$ do T_1? Pretpostavimo da čvor τ u stablu $T_{\max}$ ima konačnu djecu čvorova τ_L i τ_D. Tada vrijedi

$$R^{re}(\tau) \geq R^{re}(\tau_L) + R^{re}(\tau_D) \tag{2.16}$$

Ukoliko u prethodnom izrazu (2.16) imamo jednakost, onda skraćujemo konačne čvorove τ_L i τ_D iz stabla. Postupak ponavljamo dok više ne postoji nijedno skraćivanje, te je rezultat toga stablo T_1.

Iduće pronalazimo T_2. Neka je τ bilo koji čvor koji nije konačan od T_1, neka je T_τ podstablo čiji korijen je τ i neka je $\tilde{T}_\tau = \{\tau_1', \ldots, \tau_1 L_\tau{}'\}$ skup konačnih čvorova od T_τ. Neka je

$$R^{re}(T_\tau) = \sum_{\tau' \in \tilde{T}_\tau} R^{re}(\tau') = \sum_{l'=!}^{L_\tau} R^{re}(\tau_{l'}'), \tag{2.17}$$

tada je $R^{re}(\tau) > R^{re}(T_\tau)$. Sada, stavimo

$$R_\alpha(T_\tau) = R^{re}(T_\tau) + \alpha|\tilde{T}_\tau|. \tag{2.18}$$

Dok god je $R_\alpha(\tau) > R_\alpha(T_\tau)$, podstablo T_τ ima manju posljedičnu složenost naspram njegovog korijena τ stoga se isplati zadržati T_τ. Koristeći formule (2.12) i (2.18) dobivamo

$$\alpha < \frac{R^{re}(\tau) - R^{re}(T_\tau)}{|\tilde{T}_\tau| - 1}. \tag{2.19}$$

Desna strane prethodne nejednakosti, koja je pozitivna, izračunava smanjenje R^{re} u odnosu na povećanje broja konačnih čvorova. Za $\tau \in T_1$, definiramo

$$g_1(\tau) = \frac{R^{re(\tau) - R^{re}(T_{1,\tau})}}{|\tilde{T}_{1,\tau}| - 1} \quad \tau \notin \tilde{T}(\alpha_1), \tag{2.20}$$

gdje je $T_{1,\tau}$ jednak T_τ. Tada, $g_1(\tau)$ možemo razmatrati kao kritičnu vrijednost za α, dok god je $g_1(\tau) > \alpha_1$ ne skraćujemo prolazne čvorove $\tau \in T_1$.

Definiramo najslabiji čvor (eng. *weakest-link node*) $\tilde{\tau}_1$ kao čvor u T_1 koji zadovoljava

$$g_1(\tilde{\tau}_1) = \min_{\tau \in T_1} g_1(\tau). \tag{2.21}$$

Dok se α povećava, $\tilde{\tau}_1$ je prvi čvor za koji $R_\alpha(\tau) = R_\alpha(T_\tau)$, stoga je $\tilde{\tau}_1$ poželjan za $T_{\tilde{\tau}_1}$. Stavimo $\alpha_2 = g_1(\tilde{\tau}_1)$ i definiramo podstablo $T_2 = T(\alpha_2)$ iz T_1 skraćivanjem podstabla $T_{\tilde{\tau}_1}$ od T_1 tako da $\tilde{\tau}_1$ postane konačan čvor.

Da bismo pronašli T_3, moramo pronaći najslabiji čvor $\tilde{\tau}_2 \in T_2$ iz kritične vrijednosti

$$g_2(\tau) = \frac{R^{re(\tau) - R^{re}(T_{2,\tau})}}{|\tilde{T}_{2,\tau}| - 1}, \quad \tau \in T(\alpha_2), \tau \notin \tilde{T}(\alpha_2) \tag{2.22}$$

gdje je $T_{2,\tau}$ dio od T_τ koji je sadržan u T_2. Stavimo sada

$$\alpha_3 = g_2(\tilde{\tau}_2) = \min_{\tau \in T_2} g_2(\tau), \tag{2.23}$$

i definiramo podstablo T_3 iz T_2 skraćivanjem podstabla $T_{\tilde{\tau}_2}$ od T_2 tako da $\tilde{\tau}_2$ postane konačan čvor. Dalje radimo analogno konačan broj puta.

Kao što smo spomenuli prije, možemo imati više minimizirajućih podstabala za svaki α. Pitanje koje si postavljamo, kako izaberemo između njih? Za danu vrijednost α, $T(\alpha)$ zovemo najmanje minimizirajuće podstablo (eng. *smallest minimizing subtree*) ako je minimizirano podstablo (odnosno zadovolja 2.14) i zadovoljava sljedeći uvjet:

$$\text{ako } R_\alpha(T) = R_\alpha\left(T(\alpha)\right), \text{ onda } T \succ T(\alpha). \tag{2.24}$$

U prethodnom izrazu, $T \succ T(\alpha)$ znači da je $T(\alpha)$ podstablo od T i ima manje konačnih čvorova od T. Taj uvjet kaže da u bilo kojem slučaju $T(\alpha)$ se uzima kao najmanje stablo između svih stabala koji minimiziraju R_α. L. Breiman je u svojoj knjizi "Classificartion and Regression Tree" pokazao kako za svaki α postoji jedinstveno najmanje minimizirano podstablo.

U poglavlju 3 za određivanje klasifikacijskog stabla koristimo programski jezik SAS i ugrađenu proceduru HPSPLIT. Navedena procedura koristi najslabiju-vezu skraćivanja (eng. *weakest-link prunning*) koju je opisao Breiman, da bi se kreirao niz $\alpha_1, \alpha_2, \ldots, \alpha_m$ vrijednosti i odgovarajući niz ugniježdenih podstabala $T_1, \ldots, T_m$.

Odnosno, dobivamo konačni rastući niz parametara složenosti,

$$0 = \alpha_0 < \alpha_1 < \alpha_2 < \cdots < \alpha_M, \tag{2.25}$$

koji odgovara konačnom nizu ugniježdenih podstabala od $T_{\max}$,

$$T_{\max} = T_0 \succ T_1 \succ T_2 \succ \cdots \succ T_M, \tag{2.26}$$

gdje je $T_k = T(\alpha_k)$ jedinstveno najmanje minimizirano podstablo za $\alpha \in [\alpha_k, \alpha_{k+1})$, i T_M je podstablo koje se sastoji samo od korijena.

Počinjemo sa T_1 i povećavamo α sve dok $\alpha = \alpha_2$ odredi najslabiji čvor $\tilde{\tau}_1$, onda skratimo podstablo $T_{\tilde{\tau}_1}$ s tim čvorom tako da je $\tilde{\tau}_1$ korijen. Tako dolazimo do T_2. Ponovimo ovu proceduru pronalaženjem $\alpha = \alpha_3$ i najslabijeg čvora $\tilde{\tau}_2$ u T_2, te skratimo podstablo $T_{\tilde{\tau}_2}$ s tim čvorom kao korijen. Time dolazimo do T_3. Proces skraćivanja ponavljamo dok ne dođemo do T_M.

Pronalazak optimalnog podskupa iz navedenog niza je onda zapravo pitanje određivanja optimalne vrijednosti parametra složenosti α. To se radi koristeći validacijsko particioniranje (eng. *validation partition*) ili krosvalidaciju (eng.*cross validation*). [4, 6]

Krosvalidacija

U V-strukih krosvalidacija (eng. *V-fold cross-validation*, u daljnjem tekstu CV/V) nasumično podijelimo podatke $\mathcal{D}$ u V disjunktne podskupove podjednake veličine tako da je $\mathcal{D} = \bigcup_{v=1}^{V} \mathcal{D}_v$ gdje je $\mathcal{D}_v \cap \mathcal{D}_{v'} = \emptyset$, $v \neq v'$, te je uobičajena veličina V od 5 do 10. Zatim napravimo V različitih skupova iz $\{\mathcal{D}_v\}$ koristeći $\mathcal{L}_v = \mathcal{D} - \mathcal{D}_v$ kao v-ti skup podataka za treniranje i $\mathcal{T}_v = \mathcal{D}_v$ kao v-ti skup testnih podataka, za $v = 1, 2, \ldots, V$. Ako svaki $\{\mathcal{D}_v\}$ ima isti broj observacija, onda svaki skup podataka za treniranje će biti $\left(\frac{V-1}{V}\right) \times 100$ postotak originalnog skupa podataka.

Napravimo v-to stablo $T_{\max}^{(v)}$ koristeći v-ti skup podataka za treniranje $\mathcal{L}_v, v = 1, 2, \ldots, V$, te fiksiramo vrijednost parametara složenosti α. Neka je $T^{(v)}(\alpha)$ najbolje

skraćeno podstablo od $T_{\max}^{(v)}, v = 1, 2, \ldots, V$. Sada, svaku observaciju u v-tom skupu testiranih podataka $\mathcal{T}_v$ provučemo kroz stablo $T_{\max}^{(v)}, v = 1, 2, \ldots, V$. Neka $N_{ij}^{(v)}(\alpha)$ označava broj j-te grupe observacija testiranih podataka $\mathcal{T}_v$ koje su grupirane kao da su iz i-te grupe, za $i, j = 1, 2, \ldots, K, v = 1, 2, \ldots, V$. Kako je $\mathcal{D} = \bigcup_{v=1}^{V} \mathcal{T}_v$ disjunktna suma, tada je ukupan broj observacija iz j-ih grupa koje su grupirane kao da su iz i-te jednak $n_{ij}(\alpha) = \sum_{v=1}^{V} n_{ij}^{(v)}(\alpha), i, j = 1, 2, \ldots, K$. Ako stavimo da je n_j broj observacija u $\mathcal{D}$ koje pripadaju u j-tu grupu, $j = 1, 2, \ldots, K$ i pretpostavimo da je pogreška grupiranja jednaka za sve grupe, tada za dani α imamo da je

$$R^{CV/V}\left(T(\alpha)\right) = n^{-1} \sum_{i=1}^{K} \sum_{j=1}^{K} n_{ij}(\alpha) \tag{2.27}$$

procijenjena greška grupiranja nad $\mathcal{D}$, gdje $T(\alpha)$ minimizira podstablo $T_{\max}$.

Posljednji korak je pronaći podstablo prave veličine. Breiman je predložio ocjenjivanje formule (2.19) sa nizom vrijednosti $\alpha'_k = \sqrt{\alpha_k \alpha_{k-1}}$, gdje je α'_k geometrijska sredina intervala $[\alpha_k, \alpha_{k+1})$ u kojem je $T(\alpha) = T_k$. Stavimo

$$R^{CV/V}(T_k) = R^{CV/V}\left(T(\alpha'_k)\right). \tag{2.28}$$

Onda, izaberemo najbolje skraćeno podstablo T_* po pravilu:

$$R^{CV/V}(T_*) = \min_k R^{CV/V}(T_k) \tag{2.29}$$

i koristimo $R^{CV/V}(T_*)$ kao njegovu procijenjenu grešku grupiranja. [4]

Poglavlje 3

Prepoznavanje i analiza spola prema akustičnim karakteristikama glasa i govora

3.1 Pojmovnik

- N = broj observacija
- Mean = aritmetička sredina
- Std Dev = standardna devijacija
- 95 % CL Mean = 95 % pouzdani interval aritmetičke sredine
- 95 % CL Std Dev = 95 % pouzdani interval standardne devijacije
- Std Err (Standard Error) = standardna pogreška
- Distribution = distribucija
- Percent = postotak
- Diff = razlika
- DF = stupanj slobode (eng. *Degree of Freedom*)
- Parameter Estimate = procijenjeni parametar
- Pr > ChiSq = p-vrijednost
- 95 % Wald C.L. = 95 % Waldov pouzdani interval

- Estimated probability = procijenjena vjerojatnost
- Sensivity = osjetljivost
- Specificity = specifičnost
- Intercept Only = samo slobodni član
- Intercept and Covariates = slobodni član i ostale varijable
- Percent Concordant = postotak usklađenih vrijednosti
- Percent Discordant = postotak neusklađenih vrijednosti
- Percent Tied = postotak jednakih vrijednosti
- Pairs = parovi
- Odds ratio estimates = procjena omjera šansi
- Split criterion used = korišteni kriterij razdvajanja
- Pruning method = metoda skraćivanja
- Subtree Evaluation Criterion = kriterij evaluacije podstabla
- Number of Branches = broj grananja
- Maximum Tree Depth Requested = zadana maksimalna dubina stabla
- Maximum Tree Depth Achieved = ispunjena maksimalna dubina stabla
- Tree depth = dubina stabla
- Number of Leaves Before Pruning = broj listova prije skraćivanja
- Number of Leaves After Pruning = broj listova nakon skraćivanja

3.2 Opis podataka

Baza podataka je u listopadu 2016. godine preuzeta sa sljedeće stranice `https://www.kaggle.com/primaryobjects/voicegender`, te je kreirana za identifikaciju glasa po spolu, temeljeno na akustičnim svojstvima glasa i govora. Skup podataka sastoji se od 3.168 snimljenih glasnovnih uzoraka, prikupljenih od muških i ženskih sudionika. Uzorci glasa u bazi su prethodno obrađeni akustičnom analizom u programskom jeziku R koristeći seewave i tuneR pakete, s analiziranim frekvencijskim rasponom od 0 Hz do 280 Hz.

Skup podataka se sastoji od 3.168 observacija i 21 varijable, te u ovom odlomku dan je kratki opis varijabli korištenih u daljnjem modeliranju.

- **meanfreq** - srednja vrijednost frekvencije - ponderirani prosjek frekvencije po amplitudi (u kHz)
- **sd** - standardna devijacija frekvencije - standardna devijacija frekvencije ponderirana amplitudom
- **median** - medijan frekvencije, frekvencija na kojoj je signal podijeljen u dva frekvencijska intervala jednake energije (u kHz)
- **Q25** - prvi kvartil frekvencije (u kHz)
- **Q75** - treći kvartil frekvencije (u kHz)
- **IQR** - interkvartil (u kHz)
- **skew** - mjera asimetrije (eng. *skewness*)
- **kurt** - mjera spljoštenosti (eng. *kurtosis*)
- **sp_ent** - spektralna entropija (eng. *spectral entropy*); govori kolika je razlika u raspodjeli energije
- **sfm** - spektralna ravnina, tj. koeficijent tonaliteta (eng. *spectral flatness or tonality coefficient*)
- **mode** - mod frekvencije
- **centroid** - frekvencijski centar
- **meanfun** - prosjek osnovne frekvencije mjeren preko akustičnog signala
- **minfun** - minimum osnovne frekvencije mjeren preko akustičnog signala

- **maxfun** - maksimum osnovne frekvencije mjeren preko akustičnog signala
- **meandom** - prosjek dominantne frekvencije mjeren preko akustičnog signala
- **mindom** - minimum dominantne frekvencije mjeren preko akustičnog signala
- **maxdom** - maksimum dominantne frekvencije mjeren preko akustičnog signala
- **dfrange** - raspon dominantne frekvencije mjeren preko akustičnog signala
- **modindx** - izračunato kao akumulirana apsolutna razlika između susjednih mjerenja osnovne frekvencije podijeljeno s rasponom frekvencije (eng. *modulation index*)
- **label** - zavisna varijabla - muška ili ženska osoba

Statistička obrada podataka kao i grafički prikazi izrađeni su u statističkom programskom paketu SAS Studio (SAS OnDemand for Academics)[1]. Koristeći SAS Studio u bazi podataka dodali smo još jednu dihotomnu varijablu ekvivalentnu varijabli *label*. Varijabla *spol* je 0 ukoliko je glas muški, odnosno 1 ukoliko je glas ženski, te smo dotičnu varijablu uveli radi jednostavnosti i umjesto varijable *label*. Primijenit ćemo dvije različite analize i pokušati pokazati da one vode do istog zaključka. Prvo ćemo, koristeći logističku regresiju, proučiti koje od navednih varijabli utječu na raspoznavanje spola, te ćemo primijeniti klasifikacijska i regresijska stabala odlučivanja.

3.3 Desktiptivna statistika

U tablici 3.1 dana je deskriptivna statistika za kontinuirane varijable.

Tablica 3.1: Deskriptivna statistika za kontinuirane varijable (ispis iz SAS-a)

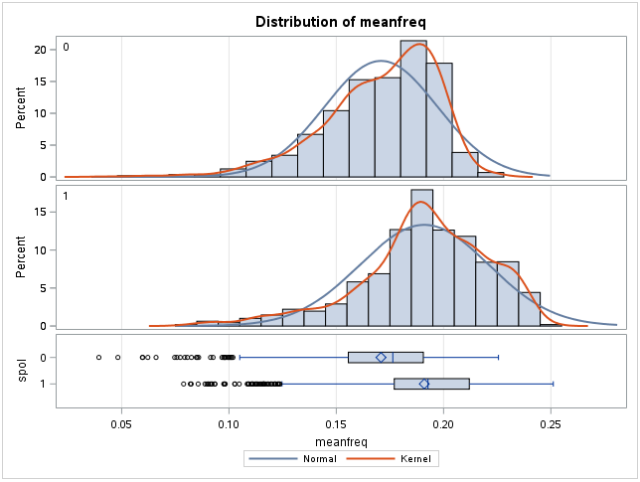
Variable	N	Mean	Std Dev	Minimum	Median	Maximum
meanfreq	3168	0.180907	0.0299178	0.0393633	0.1848384	0.251124
sd	3168	0.057126	0.0166522	0.0183632	0.0591551	0.115273
median	3168	0.185621	0.0363601	0.0109746	0.1900324	0.261225
Q25	3168	0.140456	0.0486797	0.00022876	0.1402864	0.247347
Q75	3168	0.224765	0.0236393	0.0429463	0.2256842	0.273469
IQR	3168	0.084309	0.0427831	0.0145577	0.09428	0.252225
skew	3168	3.140168	4.2405287	0.1417354	2.1971007	34.72545
kurt	3168	36.56846	134.92866	2.0684555	8.3184633	1309.61
sp_ent	3168	0.895127	0.0449795	0.7386507	0.9017668	0.981997
sfm	3168	0.408216	0.1775211	0.0368765	0.3963352	0.842936
mode	3168	0.165282	0.077203	0	0.1865986	0.28
centroid	3168	0.180907	0.0299178	0.0393633	0.1848384	0.251124
meanfun	3168	0.142807	0.0323044	0.0555653	0.1405185	0.237636
minfun	3168	0.036802	0.01922	0.0097752	0.0461095	0.204082
maxfun	3168	0.258842	0.0300773	0.1030928	0.2711864	0.279114
meandom	3168	0.829211	0.525205	0.0078125	0.7657948	2.957682
mindom	3168	0.052647	0.0632995	0.0048828	0.0234375	0.458984
maxdom	3168	5.047277	3.5211566	0.0078125	4.9921875	21.86719
dfrange	3168	4.99463	3.5200391	0	4.9453125	21.84375
modindx	3168	0.173752	0.1194544	0	0.139357	0.932374

Bitno za uočiti da sve varijable imaju jednak broj observacija što znači da u bazi nema podataka koji nedostaju (eng. *missing values*). Također iz dane tablice uočavamo da ne postoje negativni brojevi i time smo provjerili da su svi podaci pravovaljani.

Za daljnju analizu nećemo koristiti varijable *IQR*, *centroid* i *dfrange*. *Centroid* sadrži iste podatke kao i *meanfreq*, *IQR* je razlika varijabli *Q75* i *Q25*, a *dfrange* je razlika varijabli *maxdom* i *mindom*. Za preostale varijable napravit ćemo analizu po varijabli *spol*, te prikazati kako izgledaju njihove distribucije.

Tablica 3.2: Deskriptivna statistika varijable *meanfreq* po spolu (ispis iz SAS-a)

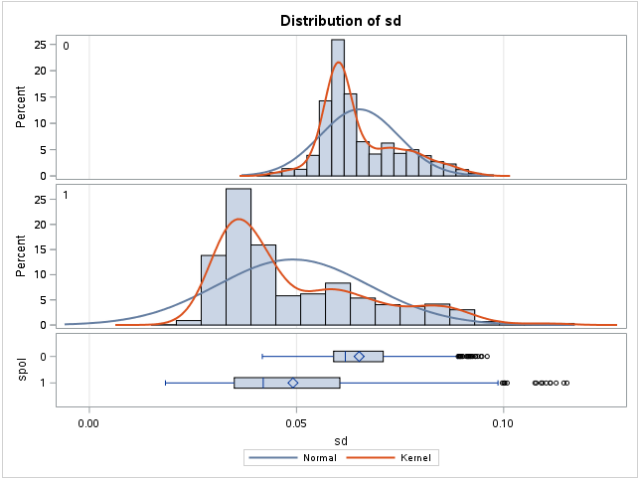
spol	N	Mean	95% CL Mean		Std Dev	95% CL Std Dev		Std Err	Minimum	Maximum
0	1584	0.1708	0.1695	0.1721	0.0263	0.0254	0.0272	0.00066	0.0394	0.2256
1	1584	0.191	0.1895	0.1925	0.03	0.029	0.031	0.000753	0.0788	0.2511
Diff		-0.0202	-0.0221	-0.0182	0.0282	0.0275	0.0289	0.001		



Slika 3.1: Distribucija varijable *meanfreq* po spolu

Tablica 3.3: Deskriptivna statistika varijable *sd* po spolu (ispis iz SAS-a)

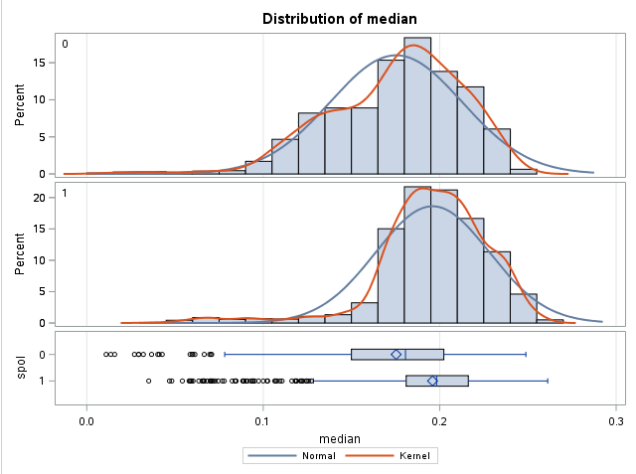
spol	N	Mean	95% CL Mean		Std Dev	95% CL Std Dev		Std Err	Minimum	Maximum
0	1584	0.0651	0.0646	0.0656	0.00945	0.00914	0.0098	0.000238	0.0417	0.096
1	1584	0.0491	0.0482	0.05	0.0184	0.0178	0.019	0.000462	0.0184	0.1153
Diff		0.016	0.015	0.017	0.0146	0.0143	0.015	0.000519		



Slika 3.2: Distribucija varijable *sd* po spolu

Tablica 3.4: Deskriptivna statistika varijable *median* po spolu (ispis iz SAS-a)

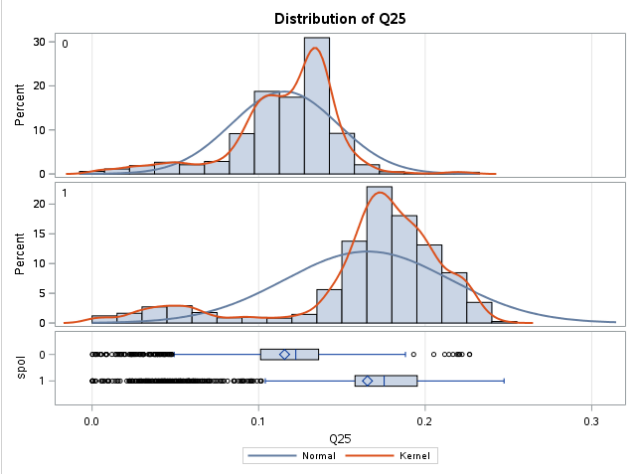
spol	N	Mean	95% CL Mean		Std Dev	95% CL Std Dev		Std Err	Minimum	Maximum
0	1584	0.1753	0.1735	0.1771	0.0374	0.0361	0.0387	0.00094	0.011	0.2488
1	1584	0.1959	0.1944	0.1975	0.0321	0.0311	0.0333	0.000808	0.0351	0.2612
Diff		-0.0206	-0.0231	-0.0182	0.0349	0.034	0.0358	0.00124		



Slika 3.3: Distribucija varijable *median* po spolu

Tablica 3.5: Deskriptivna statistika varijable *q25* po spolu (ispis iz SAS-a)

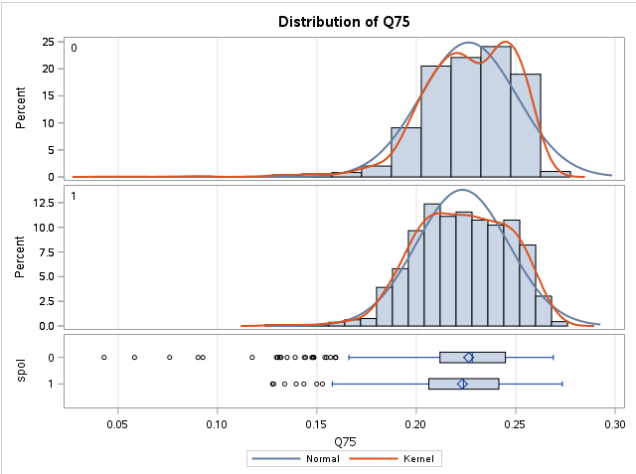
spol	N	Mean	95% CL Mean		Std Dev	95% CL Std Dev		Std Err	Minimum	Maximum
0	1584	0.1156	0.114	0.1171	0.032	0.0309	0.0332	0.000804	0.00024	0.2267
1	1584	0.1653	0.1629	0.1678	0.0498	0.0481	0.0516	0.00125	0.000229	0.2473
Diff		-0.0498	-0.0527	-0.0469	0.0418	0.0408	0.0429	0.00149		



Slika 3.4: Distribucija varijable *q25* po spolu

Tablica 3.6: Deskriptivna statistika varijable *q75* po spolu (ispis iz SAS-a)

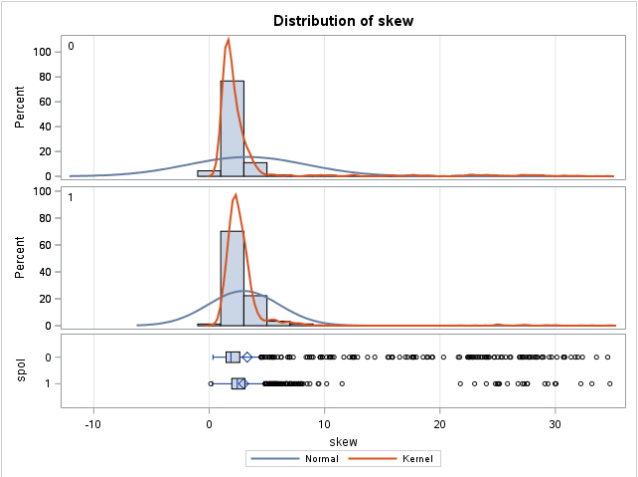
spol	N	Mean	95% CL Mean		Std Dev	95% CL Std Dev		Std Err	Minimum	Maximum
0	1584	0.2263	0.2252	0.2275	0.0241	0.0232	0.0249	0.000604	0.0429	0.2689
1	1584	0.2232	0.222	0.2243	0.0231	0.0223	0.024	0.000581	0.1276	0.2735
Diff		0.00316	0.00152	0.00481	0.0236	0.023	0.0242	0.000838		



Slika 3.5: Distribucija varijable *q75* po spolu

Tablica 3.7: Deskriptivna statistika varijable *skew* po spolu (ispis iz SAS-a)

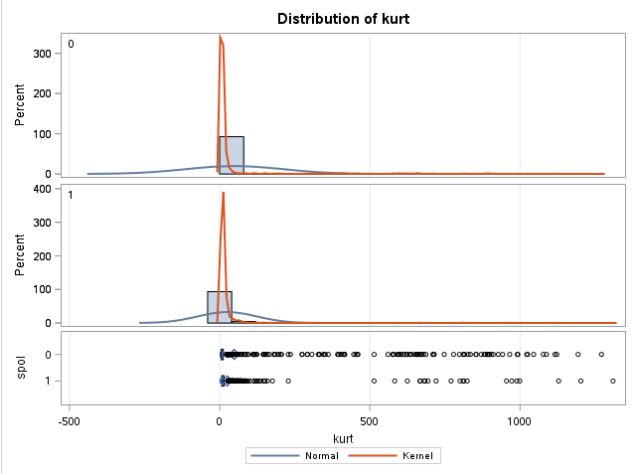
spol	N	Mean	95% CL Mean		Std Dev	95% CL Std Dev		Std Err	Minimum	Maximum
0	1584	3.2955	3.0424	3.5485	5.1352	4.9624	5.3205	0.129	0.326	34.5375
1	1584	2.9849	2.8325	3.1372	3.0915	2.9874	3.203	0.0777	0.1417	34.7255
Diff		0.3106	0.0153	0.6059	4.2384	4.1365	4.3454	0.1506		



Slika 3.6: Distribucija varijable *skew* po spolu

Tablica 3.8: Deskriptivna statistika varijable *kurt* po spolu (ispis iz SAS-a)

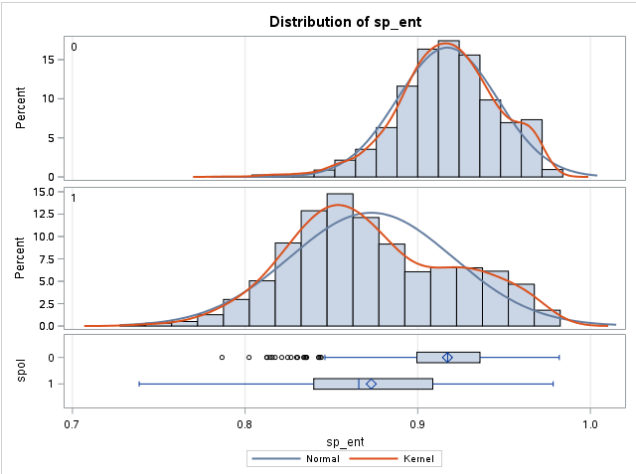
spol	N	Mean	95% CL Mean		Std Dev	95% CL Std Dev		Std Err	Minimum	Maximum
0	1584	48.3317	40.2928	56.3706	163.1	157.6	169	4.0984	2.0685	1271.4
1	1584	24.8052	19.9917	29.6187	97.6691	94.3826	101.2	2.454	2.2097	1309.6
Diff		23.5265	14.1602	32.8928	134.4	131.2	137.8	4.777		



Slika 3.7: Distribucija varijable *kurt* po spolu

Tablica 3.9: Deskriptivna statistika varijable *sp_ent* po spolu (ispis iz SAS-a)

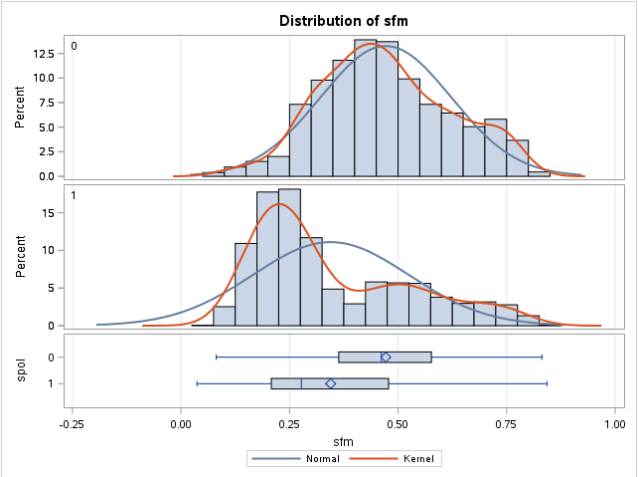
spol	N	Mean	95% CL Mean		Std Dev	95% CL Std Dev		Std Err	Minimum	Maximum
0	1584	0.9172	0.9158	0.9186	0.0289	0.028	0.03	0.000727	0.7867	0.982
1	1584	0.8731	0.8707	0.8754	0.0473	0.0457	0.049	0.00119	0.7387	0.9785
Diff		0.0441	0.0414	0.0469	0.0392	0.0383	0.0402	0.00139		



Slika 3.8: Distribucija varijable *sp_ent* po spolu

Tablica 3.10: Deskriptivna statistika varijable *sfm* po spolu (ispis iz SAS-a)

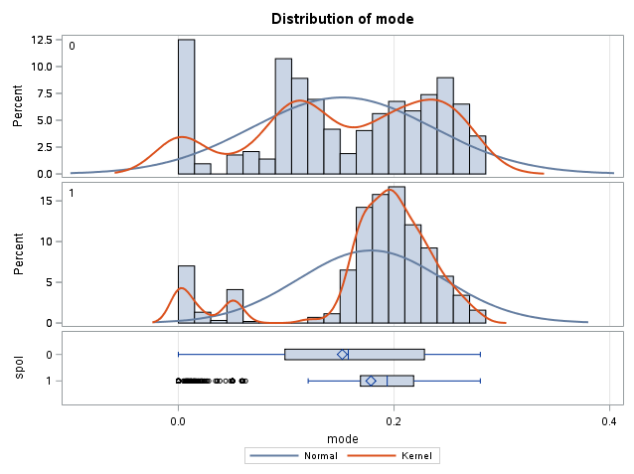
spol	N	Mean	95% CL Mean		Std Dev	95% CL Std Dev		Std Err	Minimum	Maximum
0	1584	0.4717	0.4643	0.4791	0.1505	0.1454	0.1559	0.00378	0.081	0.8313
1	1584	0.3448	0.3359	0.3536	0.1799	0.1738	0.1863	0.00452	0.0369	0.8429
Diff		0.1269	0.1154	0.1385	0.1658	0.1618	0.17	0.00589		



Slika 3.9: Distribucija varijable *sfm* po spolu

Tablica 3.11: Deskriptivna statistika varijable *mode* po spolu (ispis iz SAS-a)

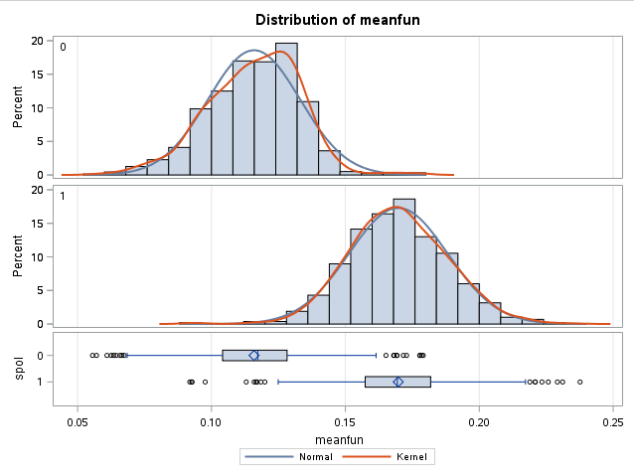
spol	N	Mean	95% CL Mean		Std Dev	95% CL Std Dev		Std Err	Minimum	Maximum
0	1584	0.152	0.1479	0.1562	0.084	0.0812	0.0871	0.00211	0	0.28
1	1584	0.1785	0.1752	0.1819	0.0672	0.0649	0.0696	0.00169	0	0.28
Diff		-0.0265	-0.0318	-0.0212	0.0761	0.0742	0.078	0.0027		



Slika 3.10: Distribucija varijable *mode* po spolu

Tablica 3.12: Deskriptivna statistika varijable *meanfun* po spolu (ispis iz SAS-a)

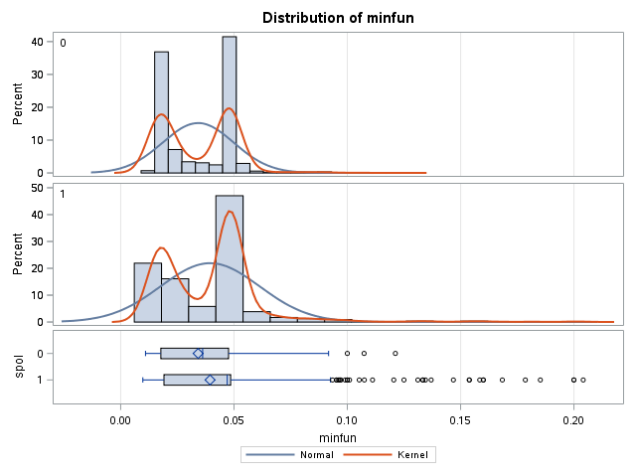
spol	N	Mean	95% CL Mean		Std Dev	95% CL Std Dev		Std Err	Minimum	Maximum
0	1584	0.1159	0.115	0.1167	0.0172	0.0166	0.0178	0.000432	0.0556	0.1791
1	1584	0.1697	0.1688	0.1707	0.0185	0.0178	0.0191	0.000464	0.0919	0.2376
Diff		-0.0539	-0.0551	-0.0526	0.0178	0.0174	0.0183	0.000634		



Slika 3.11: Distribucija varijable *meanfun* po spolu

Tablica 3.13: Deskriptivna statistika varijable *minfun* po spolu (ispis iz SAS-a)

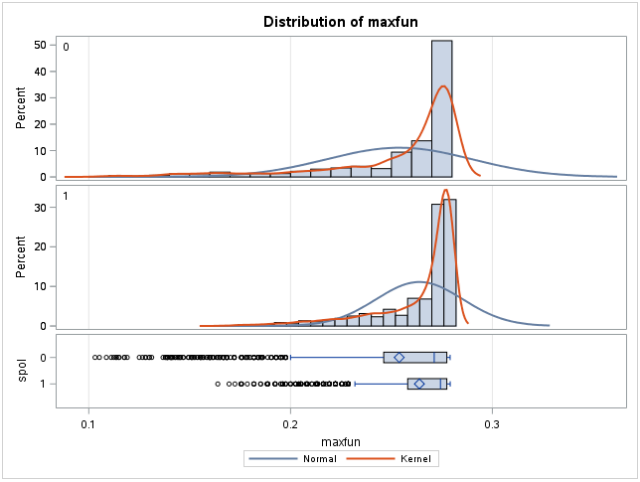
spol	N	Mean	95% CL Mean		Std Dev	95% CL Std Dev		Std Err	Minimum	Maximum
0	1584	0.0342	0.0334	0.035	0.0157	0.0152	0.0163	0.000396	0.011	0.1212
1	1584	0.0394	0.0384	0.0405	0.0218	0.0211	0.0226	0.000549	0.00978	0.2041
Diff		-0.00525	-0.00658	-0.00393	0.019	0.0186	0.0195	0.000677		



Slika 3.12: Distribucija varijable *minfun* po spolu

Tablica 3.14: Deskriptivna statistika varijable *maxfun* po spolu (ispis iz SAS-a)

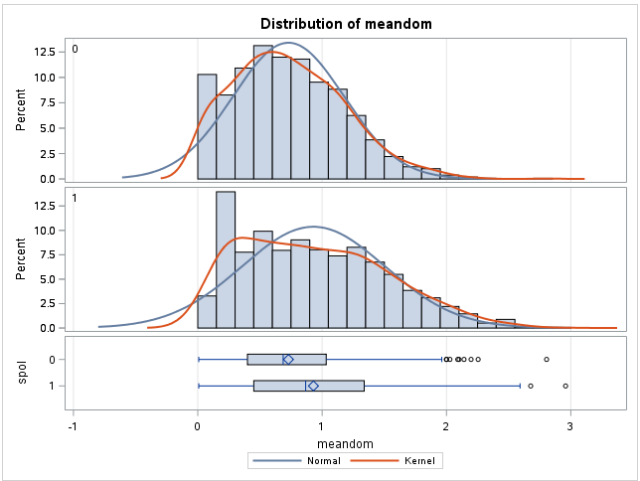
spol	N	Mean	95% CL Mean		Std Dev	95% CL Std Dev		Std Err	Minimum	Maximum
0	1584	0.2538	0.2521	0.2556	0.036	0.0348	0.0373	0.000905	0.1031	0.2791
1	1584	0.2638	0.2628	0.2649	0.0215	0.0208	0.0223	0.000541	0.1639	0.2791
Diff		-0.01	-0.0121	-0.00795	0.0297	0.0289	0.0304	0.00105		



Slika 3.13: Distribucija varijable *maxfun* po spolu

Tablica 3.15: Deskriptivna statistika varijable *meandom* po spolu (ispis iz SAS-a)

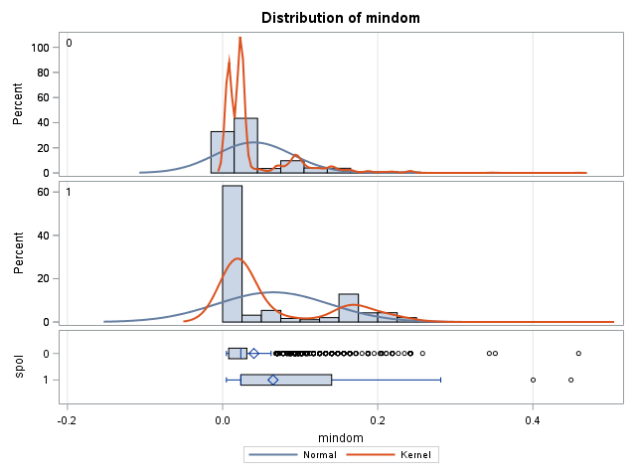
spol	N	Mean	95% CL Mean		Std Dev	95% CL Std Dev		Std Err	Minimum	Maximum
0	1584	0.7289	0.7069	0.7509	0.446	0.431	0.4621	0.0112	0.00781	2.8052
1	1584	0.9295	0.9011	0.958	0.5769	0.5575	0.5977	0.0145	0.00781	2.9577
Diff		-0.2007	-0.2366	-0.1647	0.5156	0.5032	0.5286	0.0183		



Slika 3.14: Distribucija varijable *meandom* po spolu

Tablica 3.16: Deskriptivna statistika varijable *mindom* po spolu (ispis iz SAS-a)

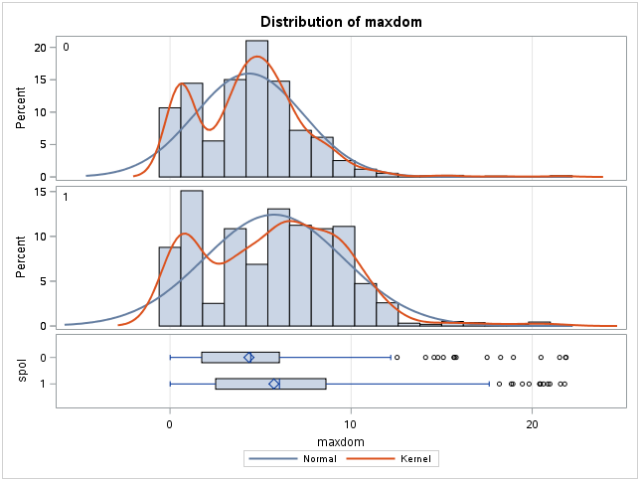
spol	N	Mean	95% CL Mean		Std Dev	95% CL Std Dev		Std Err	Minimum	Maximum
0	1584	0.0403	0.0379	0.0427	0.0492	0.0475	0.051	0.00124	0.00488	0.459
1	1584	0.065	0.0614	0.0686	0.0727	0.0703	0.0754	0.00183	0.00488	0.4492
Diff		-0.0247	-0.029	-0.0204	0.0621	0.0606	0.0637	0.00221		



Slika 3.15: Distribucija varijable *mindom* po spolu

Tablica 3.17: Deskriptivna statistika varijable *maxdom* po spolu (ispis iz SAS-a)

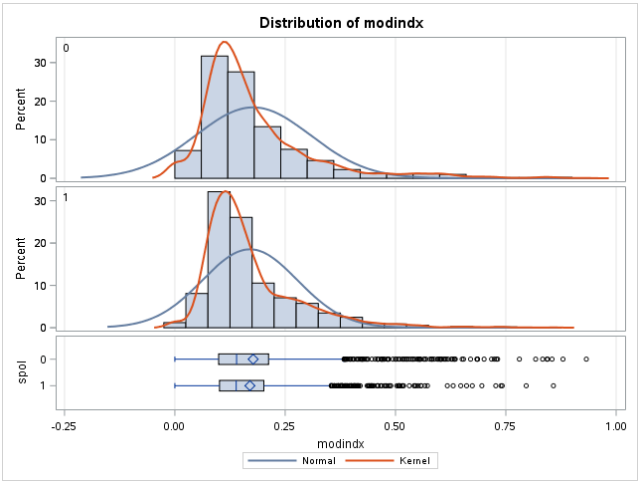
spol	N	Mean	95% CL Mean		Std Dev	95% CL Std Dev		Std Err	Minimum	Maximum
0	1584	4.3584	4.2106	4.5063	3.0003	2.8993	3.1086	0.0754	0.00781	21.8672
1	1584	5.7361	5.5462	5.926	3.854	3.7244	3.9932	0.0968	0.00781	21.7969
Diff		-1.3777	-1.6183	-1.137	3.4536	3.3706	3.5409	0.1227		



Slika 3.16: Distribucija varijable *maxdom* po spolu

Tablica 3.18: Deskriptivna statistika varijable *modindx* po spolu (ispis iz SAS-a)

spol	N	Mean	95% CL Mean		Std Dev	95% CL Std Dev		Std Err	Minimum	Maximum
0	1584	0.1774	0.171	0.1838	0.1301	0.1258	0.1348	0.00327	0	0.9324
1	1584	0.1701	0.1648	0.1754	0.1076	0.104	0.1115	0.0027	0	0.8578
Diff		0.00736	-0.00096	0.0157	0.1194	0.1165	0.1224	0.00424		



Slika 3.17: Distribucija varijable *modindx* po spolu

Nakon što smo malo proučili podatke, uočavamo da su observacije u bazi podataka podijeljenje 50:50 u odnosu na varijablu *spol*.

3.4 Univarijatna logistička regresija

U bazi podataka zavisna varijabla nam je *spol*, dok su nezavisne varijable *meanfreq, sd, median, Q25, Q75, skew, kurt, sp_ent, sfm, mode, meanfun, minfun, maxfun, meandom, mindom, maxdom* i *modindx*. Za svaku nezavisnu varijalu provodimo univarijatnu logističku regresiju, odnosno pripadni model izgleda

$$logit(p(x)) = g(x) = \ln\left[\frac{p(x)}{1-p(x)}\right] = \beta_0 + \beta_1 x.$$

Parametre β_0 i β_1 procijenjujemo metodom maksimalne vjerodostojnosti.

Za univarijatnu analizu varijable *spol* koristit ćemo naredbu PROC LOGISTIC koja je već implementirana u SAS-u, te kod za nezavisnu varijablu *meanfreq* je:

```
title" Univarijatna logisticka regresija - meanfreq";
proc logistic data=podaci_sredeni descending;
model spol=meanfreq/lackfit rsq outroc=rocgraf;
output out=crtanje predicted=prob;
run;

title"graficki prikaz p";
proc gplot data=crtanje;
plot prob*meanfreq/frame;
run;
```

Analogni kod vrijedi i za ostale varijable, samo promijenimo ime varijable za koju želimo napraviti univarijatnu logističku regresiju. Radi pravilnog rangiranja zavisne varijable potreban nam je uvjet *descending*. Općenito to znači da bi s 1 trebao biti označen *"događaj se dogodio"*, a s 0 *"događaj se nije dogodio"*. U našem slušaju, *"događaj se dogodio"* predstavlja da je pogođen ženski glas, u suprotnom muški.

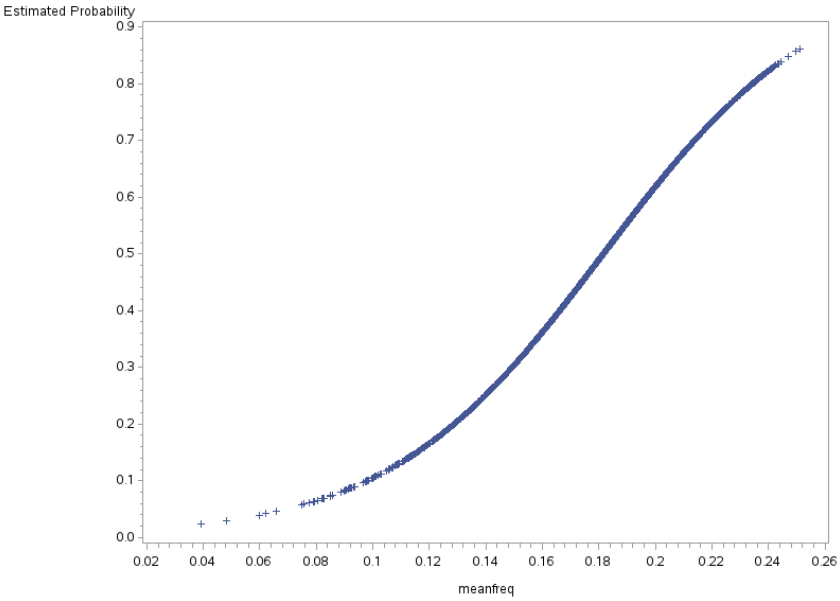
Tablica 3.19: Univarijatna logistička regresija (ispis iz SAS-a)

	DF	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	OR	95% Wald C.L.		c	-2 Log L (Intercept Only)	- 2 Log L (Intercept and Covariates)	Likelihood Ratio	Pr>ChiSq
meanfreq	1	26.2581	1.4756	316.6713	<.0001	>999.999	>999.999	>999.999	0.708	4391.781	4002.894	388.8867	<.0001
sd	1	-72.0858	2.973	587.8941	<.0001	<0.001	<0.001	<0.001	0.785	4391.781	3580.805	810.9757	<.0001
median	1	17.7073	1.1665	230.4166	<.0001	>999.999	>999.999	>999.999	0.669	4391.781	4119.787	271.9938	<.0001
Q25	1	29.3266	1.1759	622.024	<.0001	>999.999	>999.999	>999.999	0.859	4391.781	3421.578	970.2024	<.0001
Q75	1	-5.7109	1.5227	14.0667	0.0002	0.003	<0.001	0.065	0.55	4391.781	4377.537	14.2436	0.0002
skew	1	-0.0175	0.00854	4.1956	0.0405	0.983	0.966	0.999	0.35	4391.781	4387.503	4.2772	0.0386
kurt	1	-0.00144	0.000311	21.5764	<.0001	0.999	0.998	0.999	0.386	4391.781	4366.375	25.406	<.0001
sp_ent	1	-27.8353	1.1264	610.6515	<.0001	<0.001	<0.001	<0.001	0.776	4391.781	3534.852	856.929	<.0001
sfm	1	-4.4577	0.2329	366.2502	<.0001	0.012	0.007	0.018	0.725	4391.781	3966.366	425.4149	<.0001
mode	1	4.5759	0.48	90.8832	<.0001	97.116	37.907	248.805	0.581	4391.781	4297.005	94.7757	<.0001
meanfun	1	187.3	7.9719	552.0295	<.0001	>999.999	>999.999	>999.999	0.985	4391.781	936.295	3455.486	<.0001
minfun	1	15.3213	2.0383	56.4986	<.0001	>999.999	>999.999	>999.999	0.578	4391.781	4330.228	61.5529	<.0001
maxfun	1	11.9791	1.3263	81.575	<.0001	>999.999	>999.999	>999.999	0.572	4391.781	4300.492	91.2882	<.0001
meandom	1	0.7545	0.0714	111.6618	<.0001	2.127	1.849	2.446	0.594	4391.781	4274.011	117.7691	<.0001
mindom	1	6.5728	0.6164	113.6855	<.0001	715.349	213.696	>999.999	0.592	4391.781	4267.367	124.4136	<.0001
maxdom	1	0.1171	0.0109	115.9817	<.0001	1.124	1.101	1.148	0.618	4391.781	4267.205	124.5753	<.0001
modindx	1	-0.5172	0.2988	2.9971	0.0834	0.596	0.332	1.071	0.503	4391.781	4388.771	3.0097	0.0828

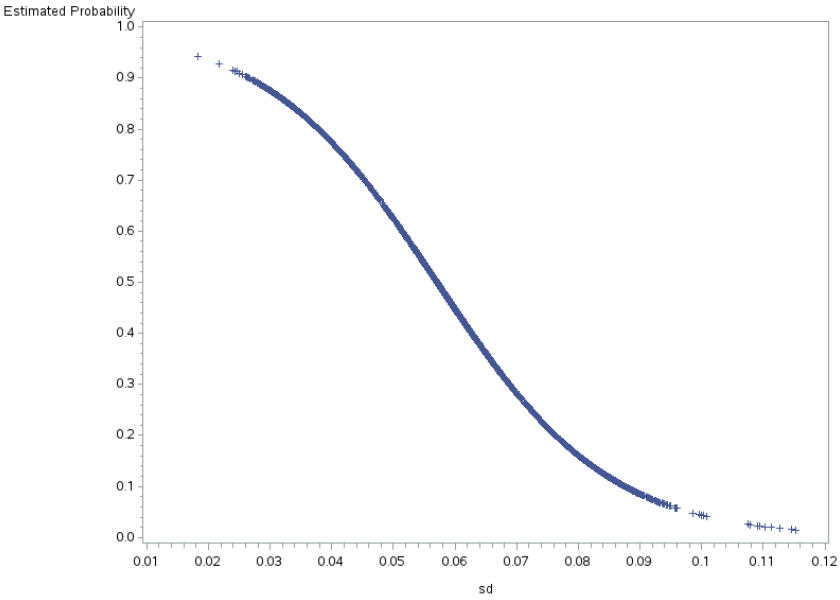
Bitno je za napomenuti da su svi modeli uspješno iskonvergirali. Iz tablice 3.19 gledajući p-vrijednosti uočavamo da su gotovo sve varijable statistički značajne, osim varijable *modindx* na razini značajnosti od 5%. Također još jedan kriterij za određivanje statistički značajnih varijabli je 95% pouzdani interval. Znači, ukoliko 95% pouzdani interval ne sadržava jedinicu, onda je varijabla statistički značajna.

Sve promatrane varijable osim *modindx* su univarijatno statistički značajni prediktori spola. Prema procijenjenom parametru povećanje *meanfun*-a statistički značajno se povećava omjer šanse da se radi o ženskom glasu. Pogledamo li npr. varijablu *sd* njen procijenjeni parametar ima negativan predznak što nam ukazuje na to da muškarci imaju puno veću varijabilnost varijable *sd* tj. da im je standardna devijacija frekvencije statistički značajno veća.

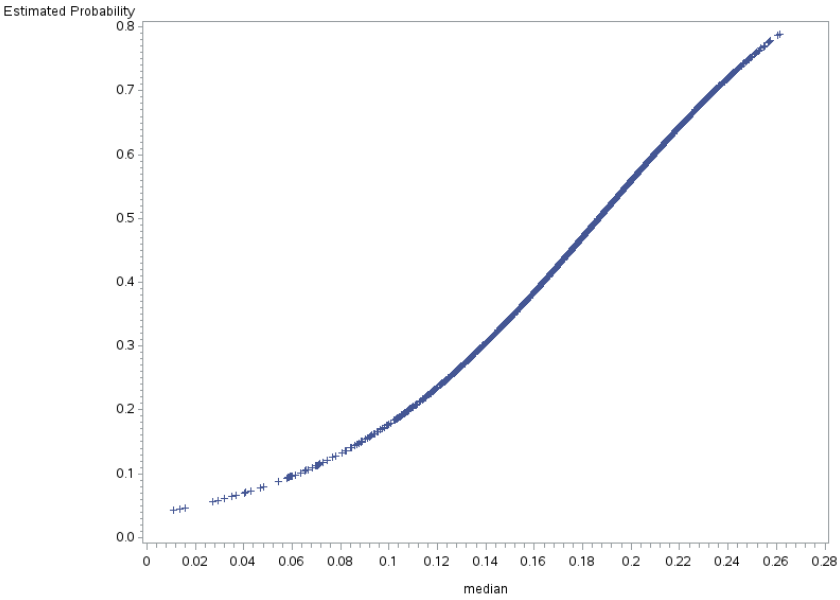
Slijedi grafički prikaz vjerojatnosti (tzv. S-krivulje) prepoznavanja ženskog glasa za svaku varijablu. Iz danih grafova možemo pročitati vrijednosti varijabli s kojim postotkom možemo prepoznati ženski glas.



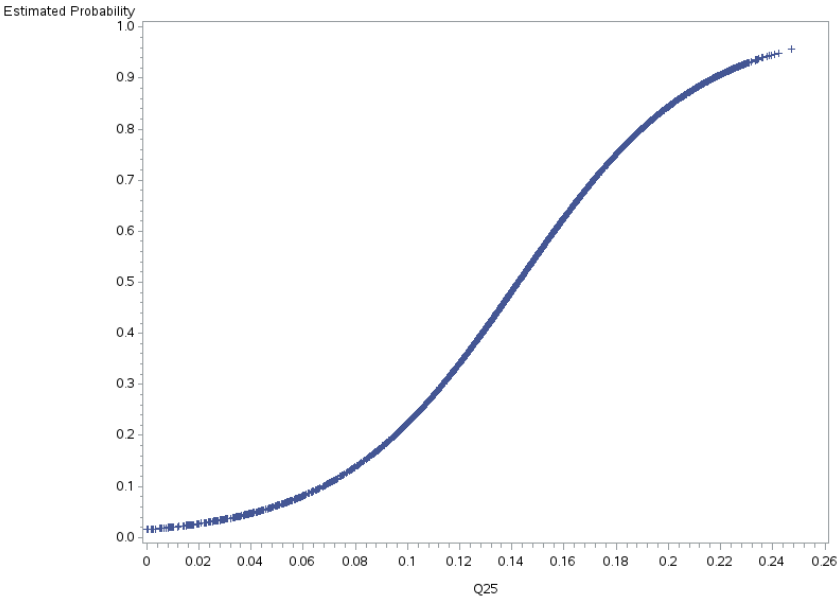
Slika 3.18: Vjerojatnost prepoznavanja ženskog glasa za varijablu *meanfreq*



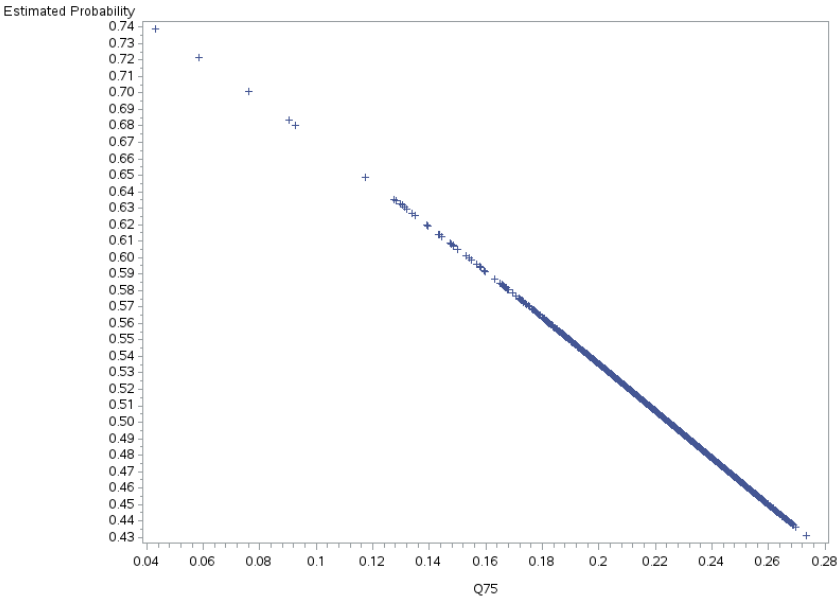
Slika 3.19: Vjerojatnost prepoznavanja ženskog glasa za varijablu *sd*



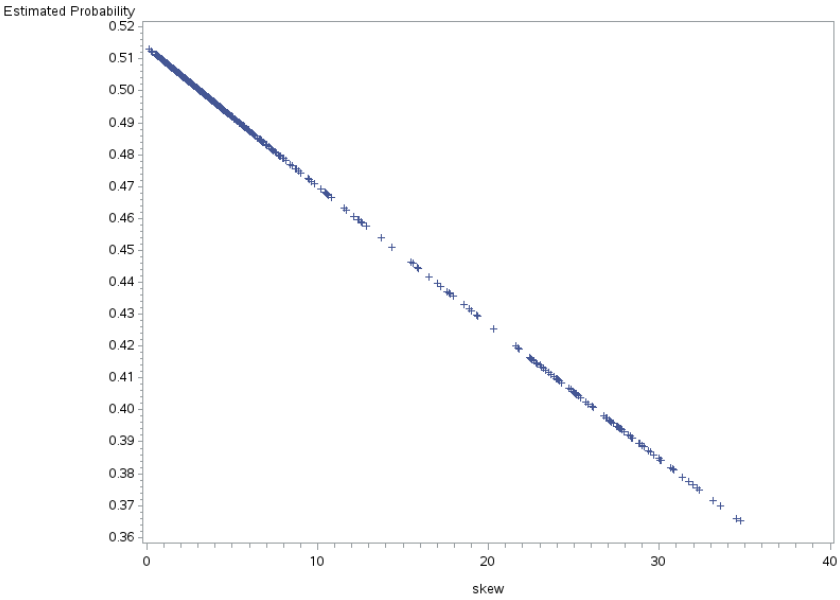
Slika 3.20: Vjerojatnost prepoznavanja ženskog glasa za varijablu *median*



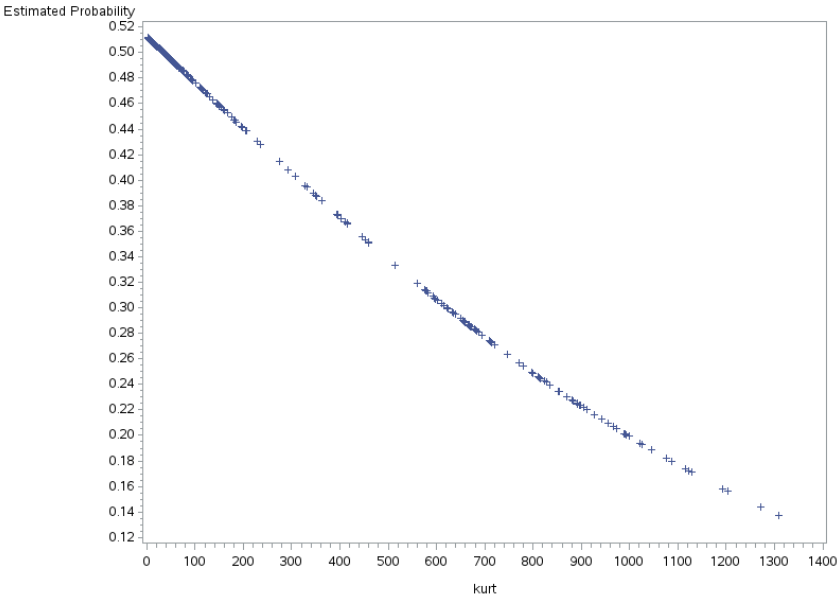
Slika 3.21: Vjerojatnost prepoznavanja ženskog glasa za varijablu *Q25*



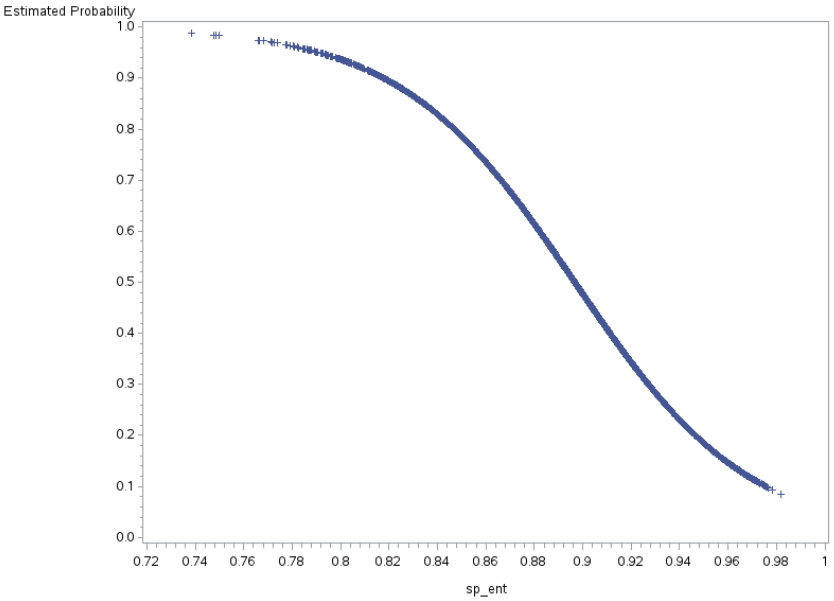
Slika 3.22: Vjerojatnost prepoznavanja ženskog glasa za varijablu *Q75*



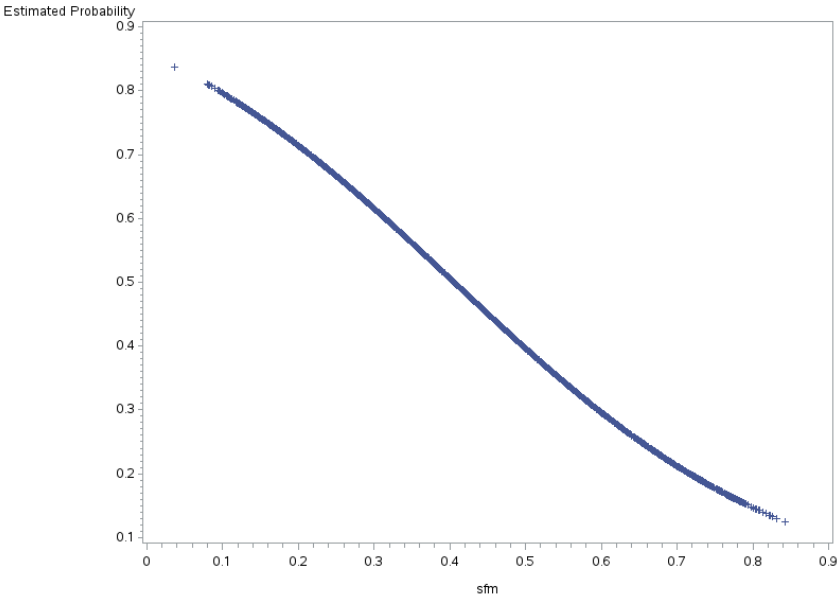
Slika 3.23: Vjerojatnost prepoznavanja ženskog glasa za varijablu *skew*



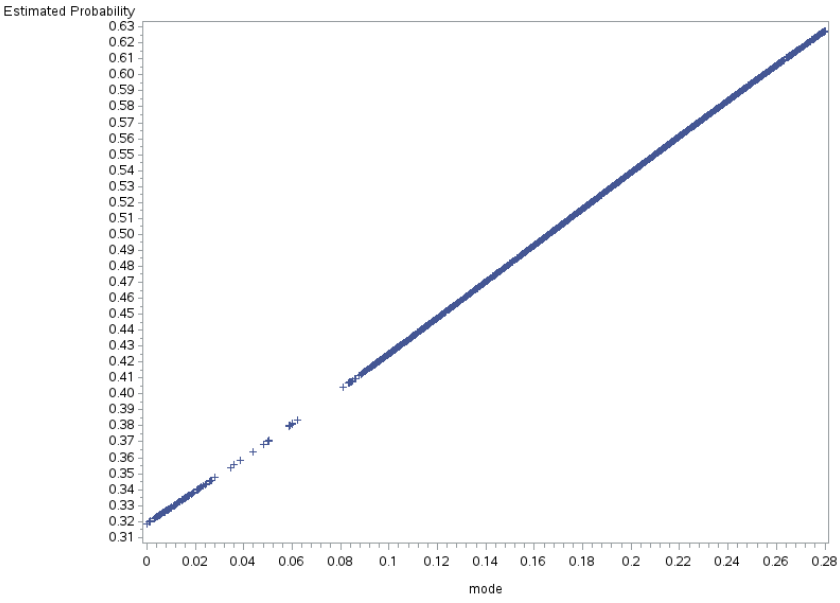
Slika 3.24: Vjerojatnost prepoznavanja ženskog glasa za varijablu *kurt*



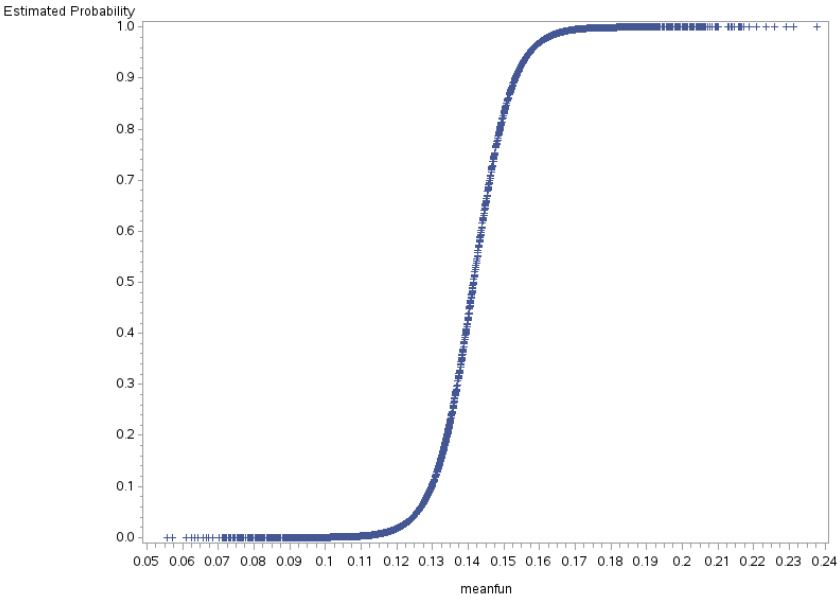
Slika 3.25: Vjerojatnost prepoznavanja ženskog glasa za varijablu *sp_ent*



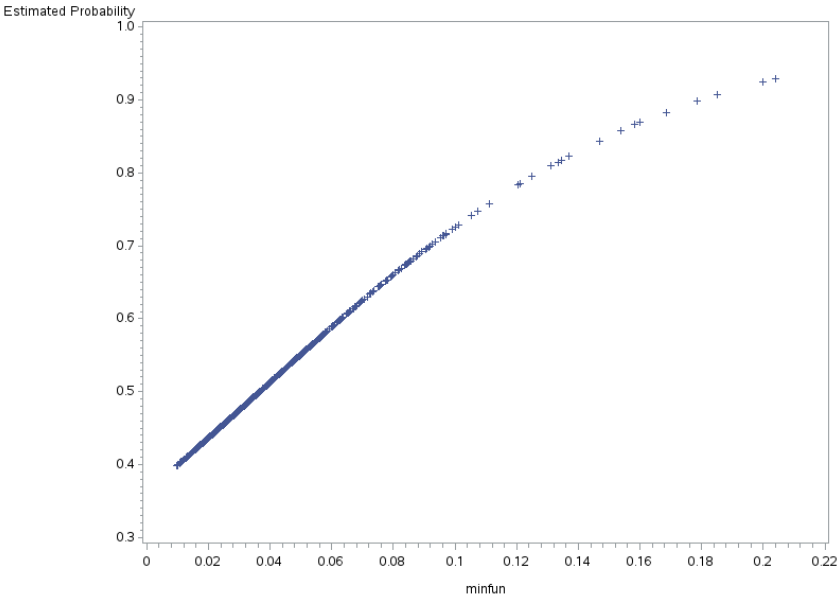
Slika 3.26: Vjerojatnost prepoznavanja ženskog glasa za varijablu *sfm*



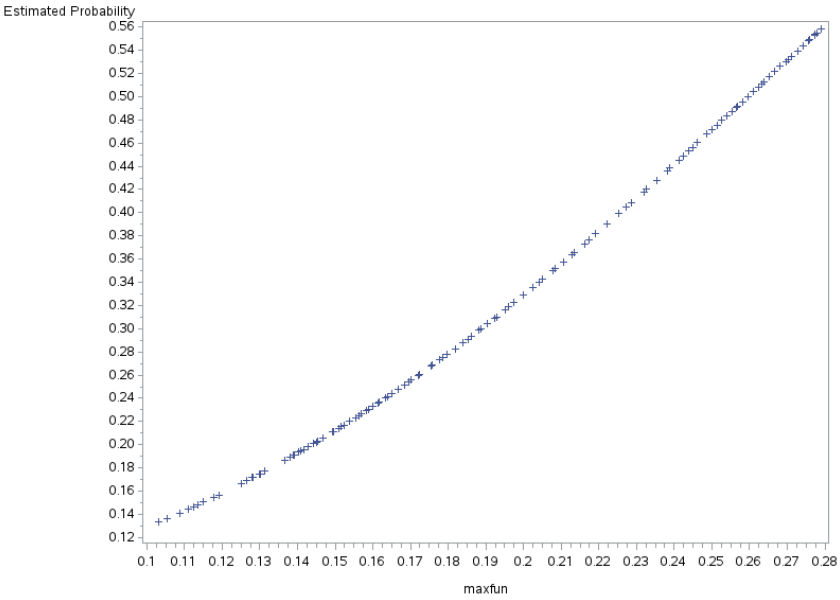
Slika 3.27: Vjerojatnost prepoznavanja ženskog glasa za varijablu *mode*



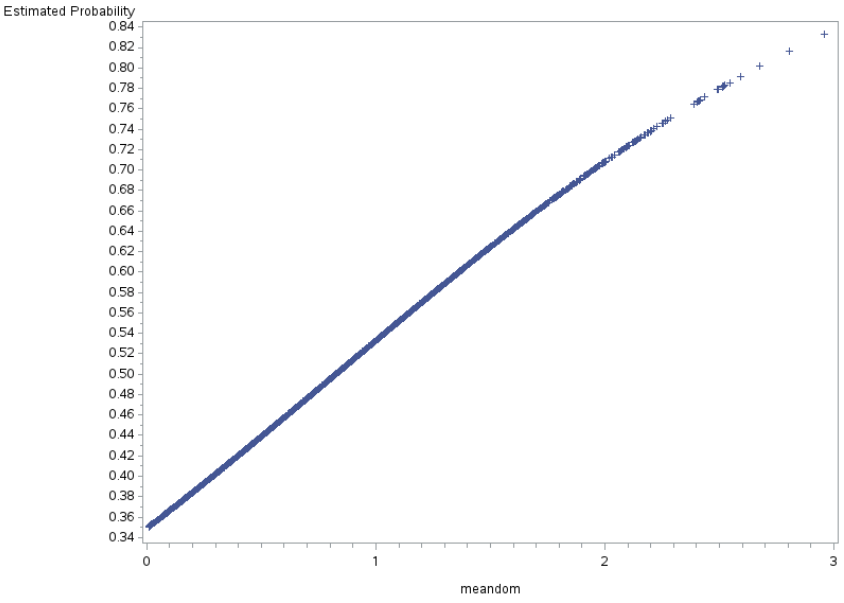
Slika 3.28: Vjerojatnost prepoznavanja ženskog glasa za varijablu *meanfun*



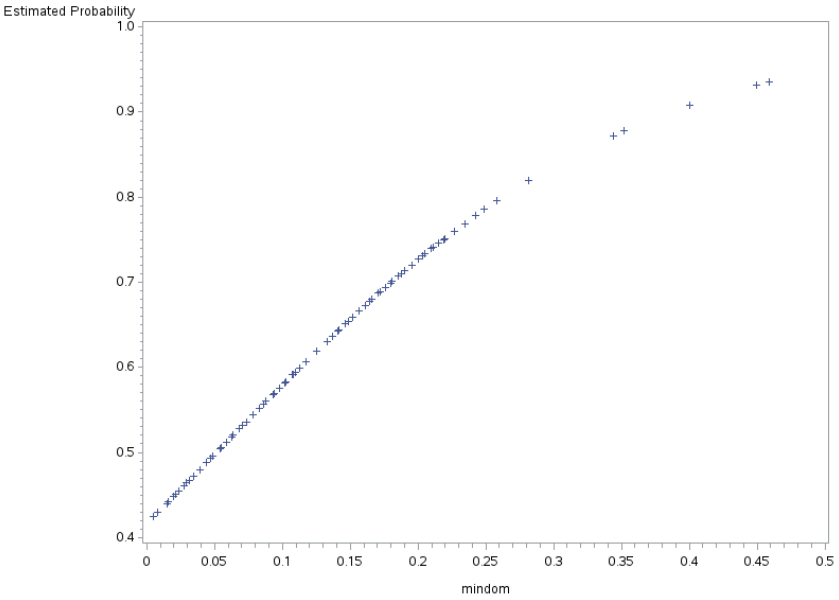
Slika 3.29: Vjerojatnost prepoznavanja ženskog glasa za varijablu *minfun*



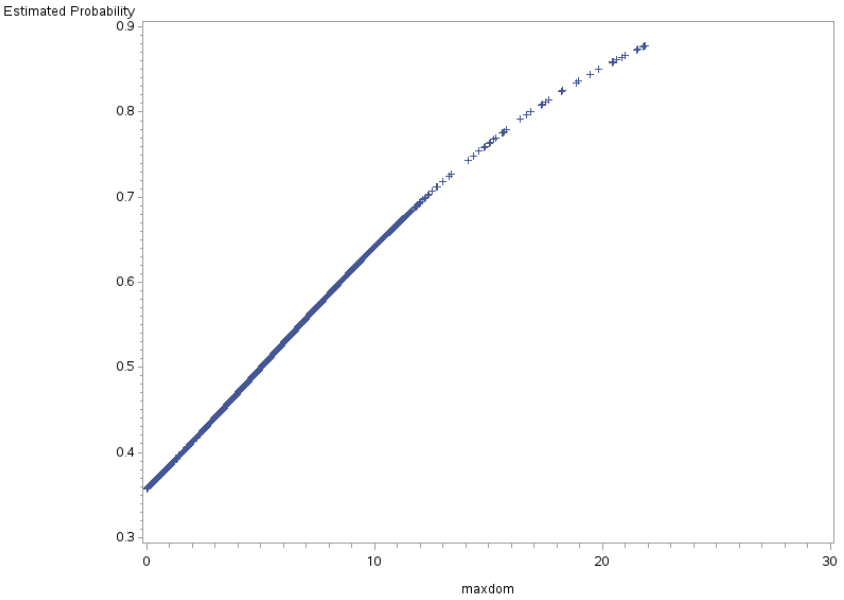
Slika 3.30: Vjerojatnost prepoznavanja ženskog glasa za varijablu *maxfun*



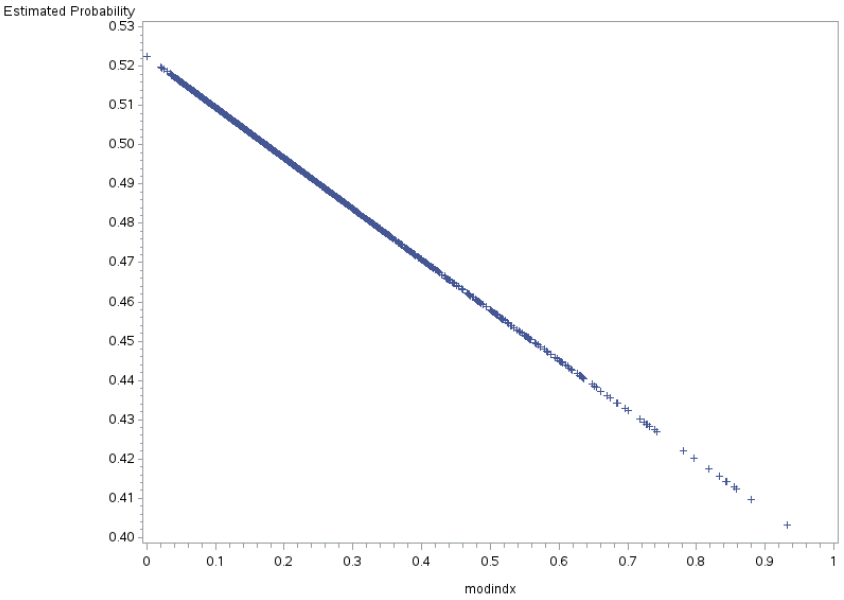
Slika 3.31: Vjerojatnost prepoznavanja ženskog glasa za varijablu *meandom*



Slika 3.32: Vjerojatnost prepoznavanja ženskog glasa za varijablu *mindom*



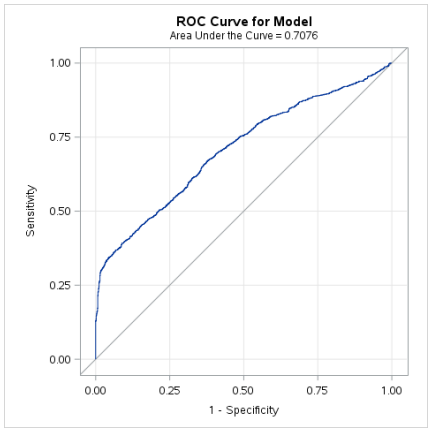
Slika 3.33: Vjerojatnost prepoznavanja ženskog glasa za varijablu *maxdom*



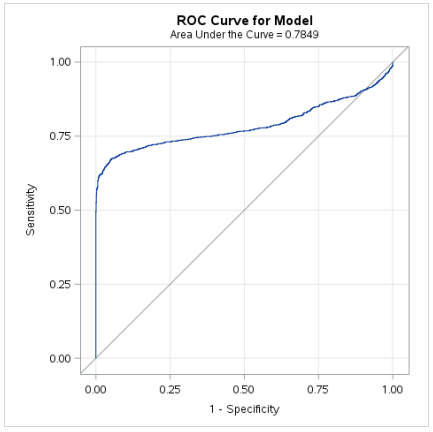
Slika 3.34: Vjerojatnost prepoznavanja ženskog glasa za varijablu *modindx*

Proučimo li vrijednosti *c*-statistike u tablici 3.19, uočavamo da varijable *Q25* ($c = 0.859$) i *meanfun* ($c = 0.985$) imaju najveću prediktivnu vrijednost. Odnosno, varijabla *meanfun* je univarijatno najznačajniji prediktor prepoznavanja ženskog glasa.

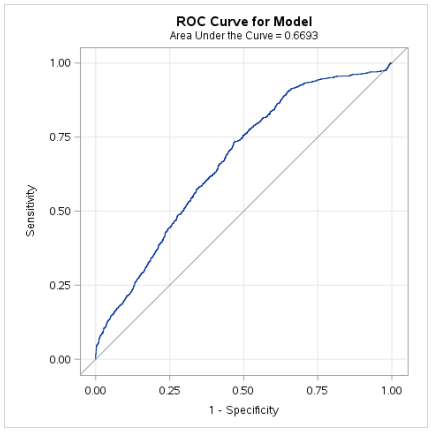
ROC krivulje za svaku varijablu prikazane su u idućim grafovima.



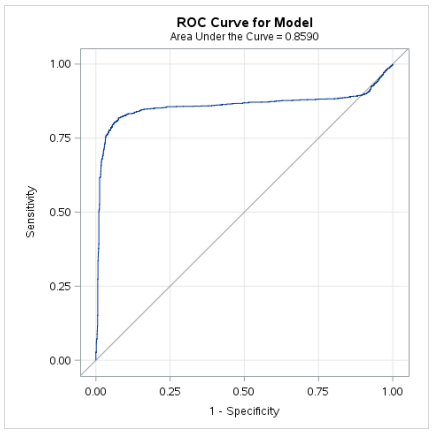
Slika 3.35: *ROC* krivulja varijable *meanfreq*



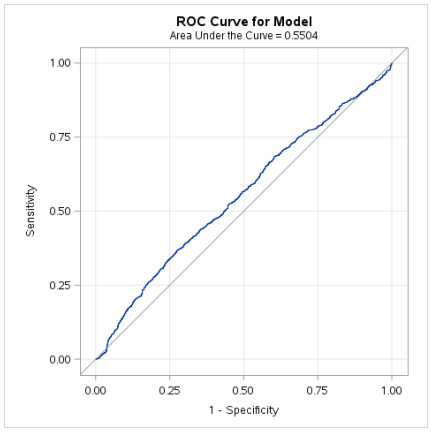
Slika 3.36: *ROC* krivulja varijable *sd*



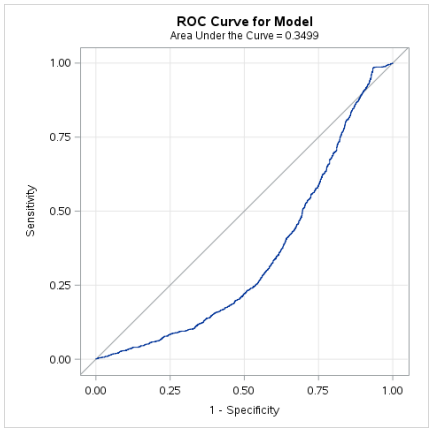
Slika 3.37: *ROC* krivulja varijable *median*



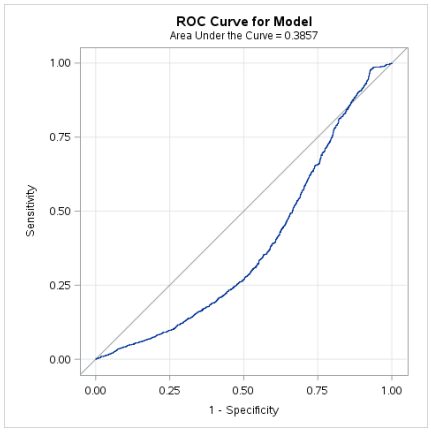
Slika 3.38: *ROC* krivulja varijable *Q25*



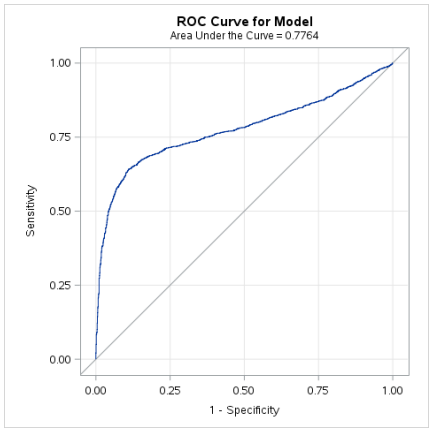
Slika 3.39: *ROC* krivulja varijable *Q75*



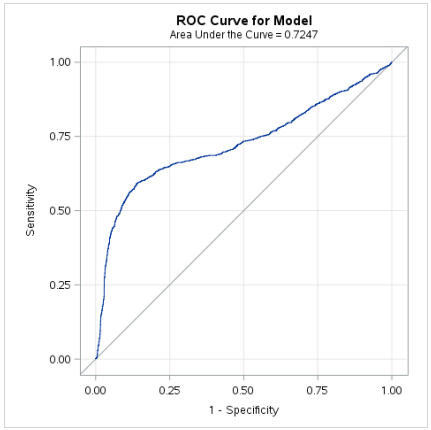
Slika 3.40: *ROC* krivulja varijable *skew*



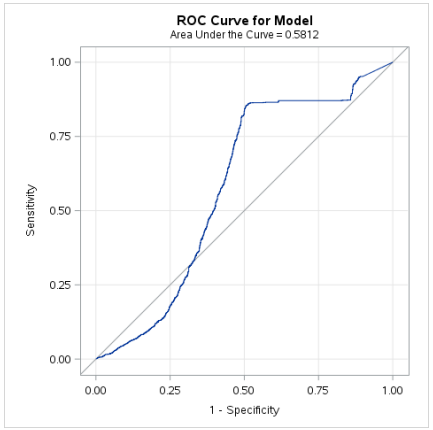
Slika 3.41: *ROC* krivulja varijable *kurt*



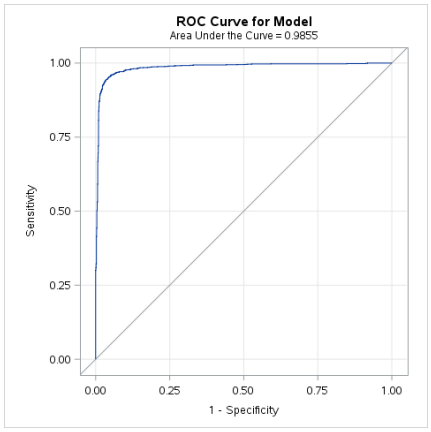
Slika 3.42: *ROC* krivulja varijable *sp_ent*



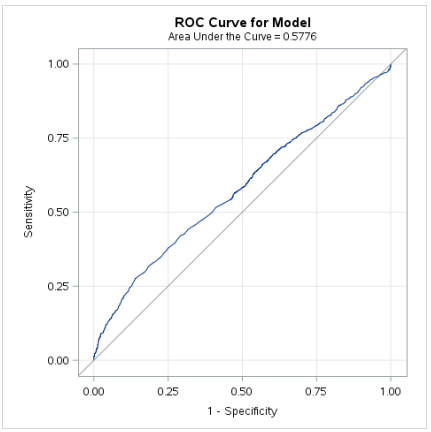
Slika 3.43: *ROC* krivulja varijable *sfm*



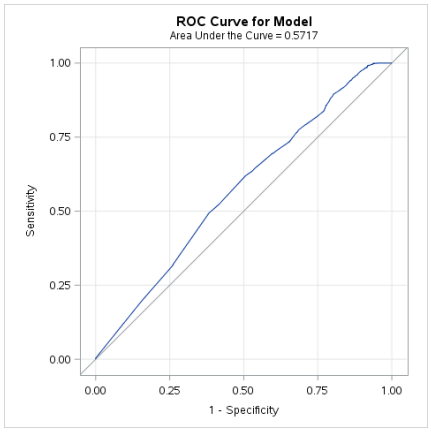
Slika 3.44: *ROC* krivulja varijable *mode*



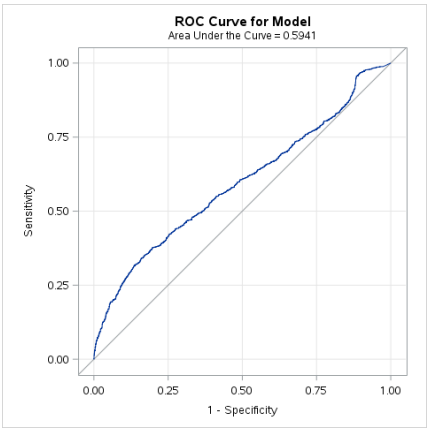
Slika 3.45: *ROC* krivulja varijable *meanfun*



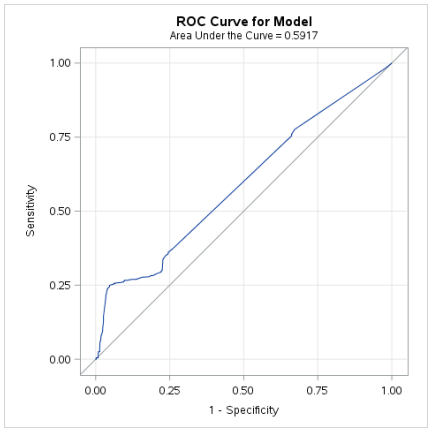
Slika 3.46: *ROC* krivulja varijable *minfun*



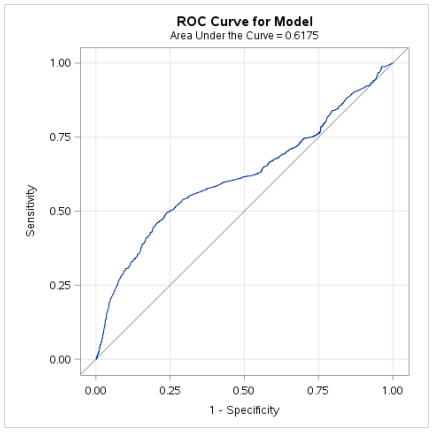
Slika 3.47: *ROC* krivulja varijable *maxfun*



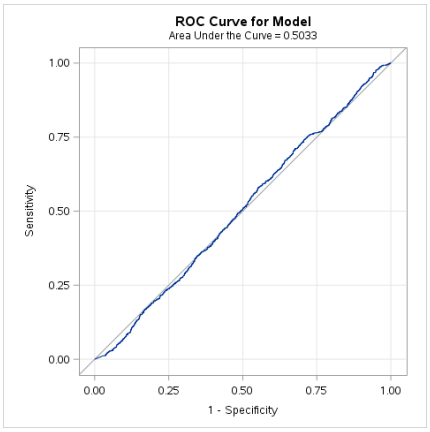
Slika 3.48: *ROC* krivulja varijable *meandom*



Slika 3.49: *ROC* krivulja varijable *min-dom*



Slika 3.50: *ROC* krivulja varijable *maxdom*



Slika 3.51: *ROC* krivulja varijable *modindx*

3.5 Multivarijatna logistička regresija

Puni model

Nakon promatrane i provedene univarijatne logističke regresije, u ovom odlomku promotriti ćemo i provesti multivarijatnu logističku regresiju. Zapravo promatramo

model oblika

$$\begin{aligned} g(x_{spol}) = \beta_0 &+ \beta_1 x_{meanfreq} + \beta_2 x_{sd} + \beta_3 x_{median} + \beta_4 x_{Q25} \\ &+ \beta_5 x_{Q75} + \beta_6 x_{skew} + \beta_7 x_{kurt} + \beta_8 x_{sp\_ent} + \beta_9 x_{sfm} \\ &+ \beta_{10} x_{mode} + \beta_{11} x_{meanfun} + \beta_{12} x_{minfun} + \beta_{13} x_{maxfun} \\ &+ \beta_{14} x_{meandom} + \beta_{15} x_{mindom} + \beta_{16} x_{maxdom} + \beta_{17} x_{modindx} \end{aligned} \tag{3.1}$$

Analogno kao i u univarijatnim modelima, želim procijeniti parametre $\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6, \beta_7, \beta_8, \beta_9, \beta_{10}, \beta_{11}, \beta_{12}, \beta_{13}, \beta_{14}, \beta_{15}, \beta_{16}$ i β_{17}. Za multivarijatnu logističku regresiju u prograskom jeziku SAS također koristimo naredbu PROC LOGISTIC, te poziv za tu naredbu izgleda:

```
proc logistic data=podaci_sredeni descending;
model spol=meanfreq sd median Q25 Q75 skew kurt
sp_ent sfm mode meanfun minfun maxfun
meandom mindom maxdom modindx/lackfit rsq outroc=rocgraf;
run;
```

Model je uspješno iskonvergirao, u suprotnom ne bi mogli promatrati takav model. Rezulati multivarijatne logističke regresije prikazani su u sljedećim tablicama.

Tablica 3.20: Statistička značajnost modela (ispis iz SAS-a)

Criterion	*Intercept Only*	*Intercept and Covariates*
- 2 Log L	4391.781	556.626

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	3835.155	17	<.0001
Score	2551.105	17	<.0001
Wald	437.6107	17	<.0001

Tablica 3.21: Procjena i statistička značajnost parametara modela (ispis iz SAS-a)

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	16.0718	9.3574	2.95	0.0859
meanfreq	1	-7.3524	46.0396	0.0255	0.8731
sd	1	-28.6291	34.7403	0.6791	0.4099
median	1	8.4755	12.8425	0.4355	0.5093
Q25	1	53.9543	11.8004	20.9054	<.0001
Q75	1	-51.8664	20.0622	6.6837	0.0097
skew	1	-0.1274	0.1693	0.5659	0.4519
kurt	1	0.00725	0.00454	2.5528	0.1101
sp_ent	1	-41.4253	10.2411	16.3621	<.0001
sfm	1	12.0302	2.5515	22.23	<.0001
mode	1	-3.2437	2.2026	2.1687	0.1408
meanfun	1	166.2	8.6716	367.3045	<.0001
minfun	1	-37.5738	9.0116	17.3847	<.0001
maxfun	1	1.3221	6.6273	0.0398	0.8419
meandom	1	-0.0701	0.4313	0.0264	0.8709
mindom	1	0.532	2.1422	0.0617	0.8039
maxdom	1	0.00461	0.0672	0.0047	0.9453
modindx	1	3.2592	1.6007	4.146	0.0417

Tablica 3.22: Omjer šansi i 95% pouzdani interval (ispis iz SAS-a)

Odds Ratio Estimates

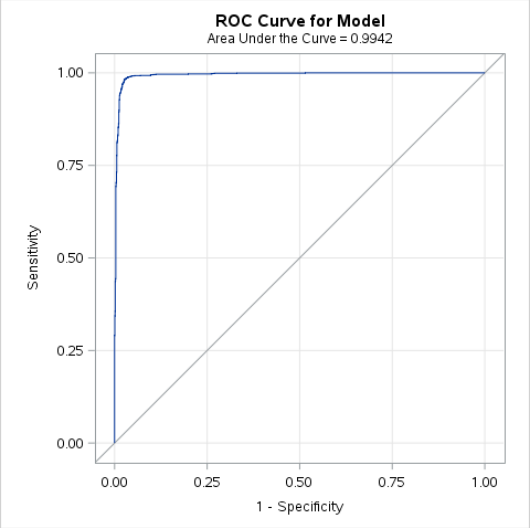
Effect	Point Estimate	95% Wald C.L.	
meanfreq	<0.001	<0.001	>999.999
sd	<0.001	<0.001	>999.999
median	>999.999	<0.001	>999.999
Q25	>999.999	>999.999	>999.999
Q75	<0.001	<0.001	<0.001
skew	0.88	0.632	1.227
kurt	1.007	0.998	1.016
sp_ent	<0.001	<0.001	<0.001
sfm	>999.999	>999.999	>999.999
mode	0.039	<0.001	2.925
meanfun	>999.999	>999.999	>999.999
minfun	<0.001	<0.001	<0.001
maxfun	3.751	<0.001	>999.999
meandom	0.932	0.4	2.171
mindom	1.702	0.026	113.361
maxdom	1.005	0.881	1.146
modindx	26.029	1.13	599.703

Tablica 3.23: Prediktivna snaga modela (ispis iz SAS-a)

Association of Predicted Probabilities and Observed Responses

Percent Concordant	99.4	**Somers' D**	0.988
Percent Discordant	0.6	**Gamma**	0.988
Percent Tied	0	**Tau-a**	0.494
Pairs	2509056	**c**	0.994

U tablici 3.20 broj stupnjeva slobode (eng. *Deegres of Freedom*, u daljnjem tekstu DF) iznosi 17 jer u modelu imamo 17 nezavisnih varijabli, također iz iste tablice za *Likelihood Ratio* i *Wald* iščitavamo p-vrijednosti manje od 5% te prema tome zaključujemo da je model statistički značajan na razini značajnosti od 5%. Iz tablice 3.21 uočavamo da su p-vrijednosti varijabli *Q25, Q75, sp_ent, sfm, meanfun, minfun* i *modindx* manje od 5%, odnosno da su statistički značajne na razini značajnosti od 5%. Isti zaključak možemo potvrditi koristeći tablicu 3.22 jer se u 95% pouzdanim intervalima ne nalazi jedinica. Iz tablice 3.23 pročitamo vrijednost c-statistike koja iznosi 0.994 i možemo zaključiti da je prediktivna snaga ovog modela jaka. Pošto smo spomenuli c-statistiku odmah uz nju prilažemo i *ROC* krivulju koja je prikazana na sljedećoj slici.



Slika 3.52: *ROC* krivulja za puni model

Vrijednosti u tablici 3.22 pod *Point Estimate* predstavljaju omjer šanse prepoznavanja ženskog glasa uz pomak vrijednosti varijabli iz modela za jednu jedinicu. Te vrijednosti su dobivene već poznatom relacijom

$$odds\ ratio(x+1, x) = e^{\beta}$$

Na primjer, za varijablu *modindx* vrijednost 26.029 dobivena je kao $e^{3.2592}$, gdje je 3.2592 procijenjeni parametar. To znači ukoliko imamo povećanje vrijednosti *modindx* za jednu jedinicu, povećava se omjer šanse za prepoznavanje ženskog glasa za približno 26 puta odnosno za približno 2500%. Za varijable *median, Q25, sfm* i *meanfun* omjer šansi je poprilično velik broj u odnosu na ostale vrijednosti, ali uzmemo li u obzir da varijabla *median* iz prijašnjih zaključaka nije statistički značajna. Možemo zaključiti da su varijable *Q25, sfm* i *meanfun* ozbiljni prediktori za prepoznavanje ženskog glasa.

Krnji model

Na temelju podataka iz prethodnog podpoglavlja zaključili smo da su varijable *Q25, sfm* i *meanfun* ozbiljni prediktori za prepoznavanje ženskog glasa, te ćemo sada napraviti multivarijatnu logističku regresiju koristeći samo navedene varijable. U stvari promatramo sljedeći model

$$g(x) = \beta_0 + \beta_1 x_{Q25} + \beta_2 x_{sfm} + \beta_3 x_{meanfun} \tag{3.2}$$

Analogno kao i prije, želimo procijeniti parametre $\beta_0, \beta_1, \beta_2$ i β_3. Za ovaj model kod u SAS izgleda:

```
proc logistic data=podaci_sredeni descending plots=roc;
model spol=Q25 sfm meanfun/lackfit rsq outroc=rocgraf;
run;
```

Model je usješno iskonvergirao, te su rezultati za ovaj model dobiveni koristeći multivarijatnu logističku regresiju prikazani su u idućim tablicama.

Tablica 3.24: Statistička značajnost modela (ispis iz SAS-a)

Criterion	*Intercept Only*	*Intercept and Covariates*
- 2 Log L	4391.781	887.584

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	3547.357	3	<.0001
Score	2228.339	3	<.0001
Wald	528.0475	3	<.0001

Tablica 3.25: Procjena i statistička značajnost parametara modela (ispis iz SAS-a)

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-32.4061	1.5021	465.4162	<.0001
Q25	1	31.0717	3.2178	93.2423	<.0001
sfm	1	4.9813	0.7801	40.7727	<.0001
meanfun	1	183.5	8.2142	499.0656	<.0001

Tablica 3.26: Omjer šansi (ispis iz SAS-a)

Odds Ratio Estimates

Effect	Point Estimate	95% Wald C.L.	
Q25	>999.999	>999.999	>999.999
sfm	145.657	31.572	671.984
meanfun	>999.999	>999.999	>999.999

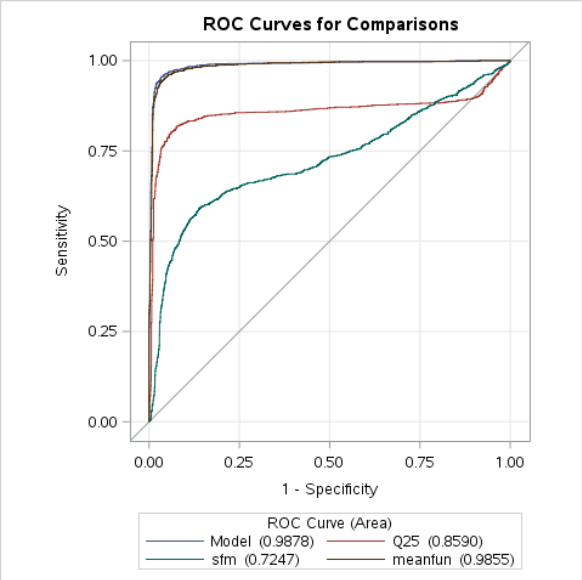
Tablica 3.27: Prediktivna snaga modela i svake varijable zasebno (ispis iz SAS-a)

ROC Association Statistics

ROC Model	Mann-Whitney				Somers' D	Gamma	Tau-a
	Area	Standard Error	95% Wald C.L.				
Model	0.9878	0.00186	0.9841	0.9914	0.9755	0.9755	0.4879
Q25	0.859	0.00797	0.8434	0.8746	0.718	0.718	0.3591
sfm	0.7247	0.0093	0.7064	0.7429	0.4493	0.4493	0.2247
meanfun	0.9855	0.00202	0.9815	0.9894	0.9709	0.9709	0.4856

Iz tablice 3.24 za *Likelihood Ratio* i *Wald* iščitavamo da su p-vrijednosti manje od 0.001 te prema tome zaključujemo da je model statistički značajan na razini značajnosti od 5%. U istoj tablici vidimo da imamo 3 stupnja slobode i to nam je neka vrste provjere da se u modelu nalaze 3 nezavisne varijable. Također, iz tablice 3.25 i pripadnih p-vrijednosti zaključujemo da su sve tri nezavisne varijable statistički značajne. Isti zaključak možemo donijeti ako promatramo 95% pouzdani interval iz tablice 3.26 pošto u sva tri slučaja jedinica se ne nalazi u intervalu. C-statistika za ovaj model sa tri nezavisne varijable iznosi 0.9878 i tu vrijednost vidimo u tablici 3.27. U istoj tablici imamo c-statistike zasebnih varijabli što smo već komentirali u poglavlju kada smo proučavali univarijatnu logističku regresiju.

C-statistika modela sa 3 nezavisne varijable je veća od c-statistike varijable *meanfreq* koja zasebno ima najveću prediktivnu snagu, pa možemo reći da ovaj model bolji od modela koji ima samo nezavisnu varijablu *meanfreq*. Na idućem grafu je prikazana usporedba *ROC* krivulja za dani model i njegove varijable zasebno.



Slika 3.53: Usporedba *ROC* krivulja

Uspoređujemo li c-statistiku za puni model ($c = 0.994$) i krnji model ($c = 0.9878$) vidimo da puni model ima za 0.62% veću prediktivnu vrijednost od krnjeg modela. Usporedimo li c-statistiku između krnjeg modela ($c = 0.9878$) i univarijatnog modela s varijablom *meanfun* ($c = 0.9855$) vidimo da je razlika još manja i iznosi 0.23%. Možemo zaključiti da je varijabla *meanfun* dovoljna za dobru predikciju spola.

3.6 Primjena CART analize

Cilj ovog poglavlja je koristeći CART analizu doći do sličnog zaključka do kojih smo došli putem logističke regresije. U ovom poglavlju također ćemo koristiti programski jezik SAS i proceduru HPSPLIT. Procedura HPSPLIT je procedura visokih performansi koja gradi stablo na temelju statističkih modela za klasifikaciju i regresiju. Procedura gradi klasifikacijsko stablo što modelira kategorijski odgovor i regresijsko stablo što modelira kontinuirani odgovor. Obje vrste stabala nazivaju se stabla odlučivanja zato što je model izražen kao niz if-then izjava.

SAS / STAT softver pruža različite metode regresije i klasifikacije i u usporedbi s drugim metodama, CART analiza je dobra jer se lako tumači i vizualizira, pogotovo kada je stablo maleno.

Glavne značajke procedure HPSPLIT su:

- pruža različite načine razdvajanja čvorova, uključujući kriterije bazirane na nečistoći (entropija, Gini index, rezidual sume kvadrata)
- pruža učinkovitu računalnu strategiju za generiranje podjele kandidata
- daje posljedičnu složenost u skraćivanju stabala
- podržava uporabu krosvalidacija i validaciju podataka za odabir najboljeg podstabla
- pruža različite metode za korištenje podataka koji nedostaju
- stvara dijagramsko stablo, graf za analizu složenosti troškova i grafove za *ROC* krivulje
- računa statističke podatke za procjenu prilagodbe modela
- računa mjere za značajne varijable
- stvara datoteku koja sadrži pravila za čvorove
- stvara izlazni skup podatka s dodjelama listova i predviđenim vrijednostima za observacije

Prije korištenja procedure HPSPLIT podatke trebamo sortirati po spolu. Slijedi prikaz koda kako smo sortili podatke te kako smo pozvali proceduru HPSPLIT.

```
data novo;
set podaci_sredeni;
proc sort; by spol;

ods graphics on;
proc hpsplit seed=15531;
class spol;
model spol (event = '1') = meanfreq sd median Q25 Q75 skew
kurt sp_ent sfm mode meanfun minfun maxfun meandom
mindom maxdom modindx;
grow entropy;
```

```
prune costcomplexity;
run;
```

U kodu smo stavili *class spol* i time smo varijablu *spol* označili kao kategorijsku varijablu odaziva za koju zahtjevamo klasifikacijsko stablo. Ostale varijable su kontinuirane (kao što smo ih opisali u poglavlju 3.2) jer nisu stavljene kod naredbe *class*. Naredbe *qrow* i *prune* kontroliraju dva osnovna aspekta za izgradnju klasifikacijskog i regresijskog stabla: rast (eng. *growing*) i skraćivanje (eng. *pruning*). Naredbu *grow* koristimo kako bi specificirali kriterij za rast stabla za rekurzivno dijeljenje roditeljskih čvorova na dječje čvorove. Zadana vrijednost za naredbu *grow* je *entropy*. Zadano je, stablo raste dok stablo ne poprimi maksimalnu dubinu rasta 10, naravno taj broj možemo promijeniti koristeći MAXDEPTH=opcija. Rezultat je često veliko stablo koje ne daje dobre prediktivne podatke. Strategija koja se preporuča da bi se izbjegli ovi problemi je skratiti stablo na neko manje podstablo koje minimizira prediktivnu grešku. Upravo za navedeno možemo koristiti naredbu *prune* koja specificira metodu skraćivanja. Zadana vrijednost je *cost-complexity*. U slučaju binarnog ishoda, EVENT=opcija se koristi za eksplicitno upravljanje razinom varijable odaziva koja predstavlja događaj od interesa za računanje područja ispod krivulje (AUC), osjetljivost, specifičnost i vrijednost *ROC* krivulje.

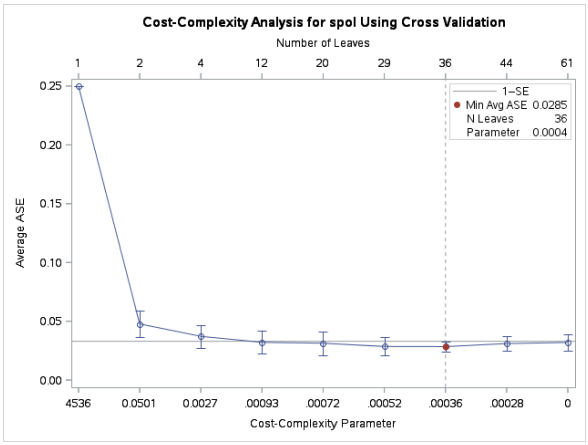
U tablici 3.28 prikazane su informacije o modeli i metode koje se koriste za rast i skraćivanja stabla.

Tablica 3.28: Informacije o modelu (ispis iz SAS-a)

Model Information	
Split Criterion Used	Entropy
Pruning Method	Cost-Complexity
Subtree Evaluation Criterion	Cost-Complexity
Number of Branches	2
Maximum Tree Depth Requested	10
Maximum Tree Depth Achieved	10
Tree Depth	10
Number of Leaves Before Pruning	68
Number of Leaves After Pruning	40
Model Event Level	0

Na slici 3.54 graf predstavlja procjenu prosječne kvadratne pogreške (ASE, eng. *Average Squared Error*) za seriju progresivno manjih podstabala koji su indek-

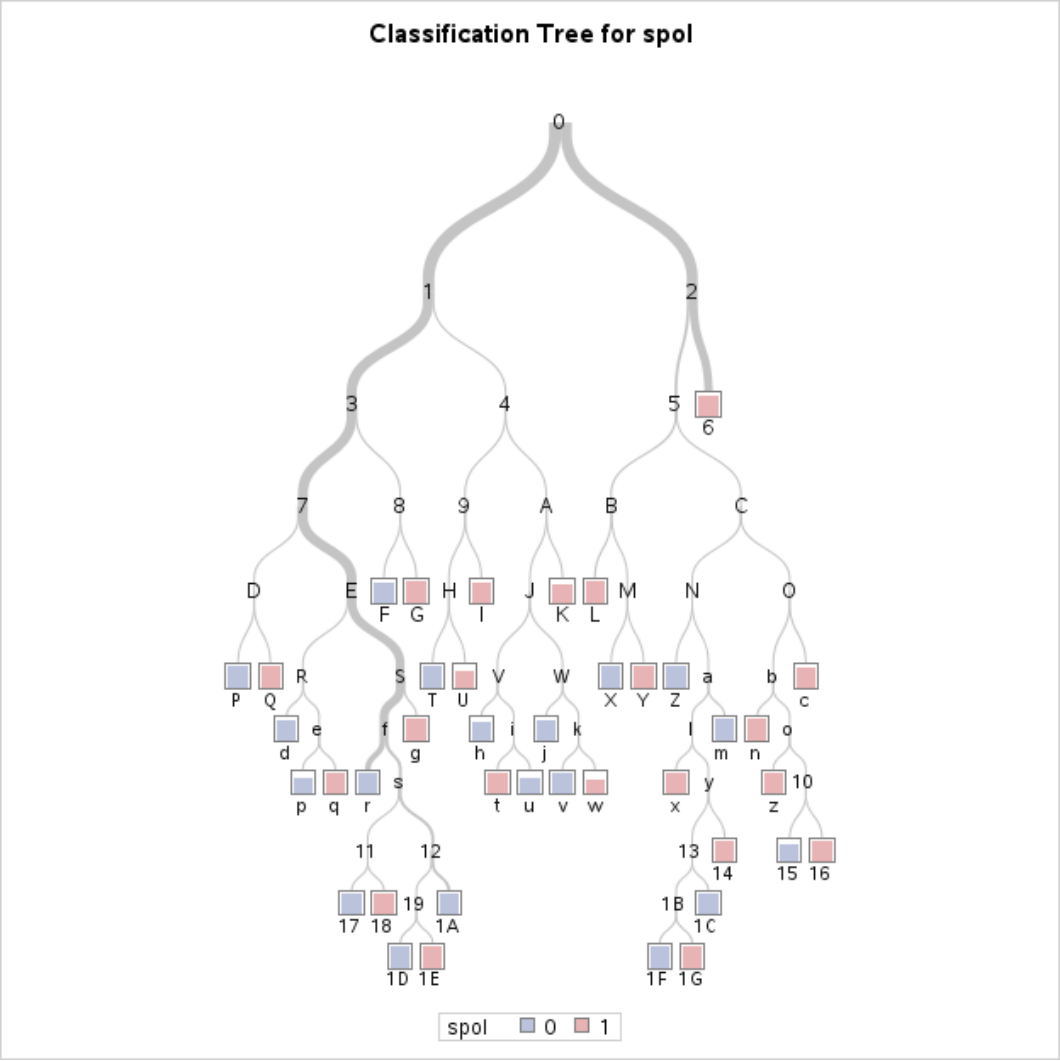
sirani parametrom složenosti troškova koji se također naziva parametar ugađanja ili skraćivanja (eng. *pruning or tuning parameter*). Graf prikazuje alat za odabir parametara koji rezultira najmanjim procijenjenim ASE. Veličina podstabla (broj listova) koji odgovara svakom parametru označen je na gornjoj vodoravnoj osi. Parametar vrijednosti 0 odgovara potpunom stablu, koji u našem slučaju ima 61 list. Prema zadanim postavkama, PROC HPSPLIT odabire parametar koji minimizira ASE, kao što pokazuje vertikalna referentna crta i točka na slici 3.54. Ovdje se minimalni ASE pojavljuje kod parametarske vrijednosti od 0.00036 što odgovara podstablu s 36 listova. Međutim, ASE za tri manja podstabla, onaj s 20 i 29 listova, ne razlikuju se od minimalnog ASE. Ovo je očigledno iz usporedbe standardnih pogrešaka za ASE, koje su prikazane vertikalnim crtama pogreške. Općenito, slika 3.54 često pokazuje odabir parametara koji odgovara manjim podstablima za koje je ASE gotovo jednak minimalnom ASE. Česti pristup za odabir parametra je pravilo "1-SE" koje odabire najmanje podstablo čiji je ASE manji od minimalnog ASE plus jedna standardna pogreška. U ovom slučaju, pravilo "1-SE" odabire podstablo s 12 listova, ali možemo vidjet da na slici njezin kružić kao da siječe liniju minimalnog ASE pa zato uzimamo podstablo s 29 listova i vrijednosti parametra 0.00052. Bitno je za napomenuti, procijenjeni ASE i njihove standarne pogreške ovise o slučajnoj raspodjeli promatranja. Dobivamo drugačije procijene ukoliko izaberemo neku drugu vrijednost SEED=vrijednost. Ukoliko želimo stablo sa 29 listova, u prijašnjem kodu trebamo promijeniti samo jednu liniju koda, *prune cost-complexity(leaves=29);*.



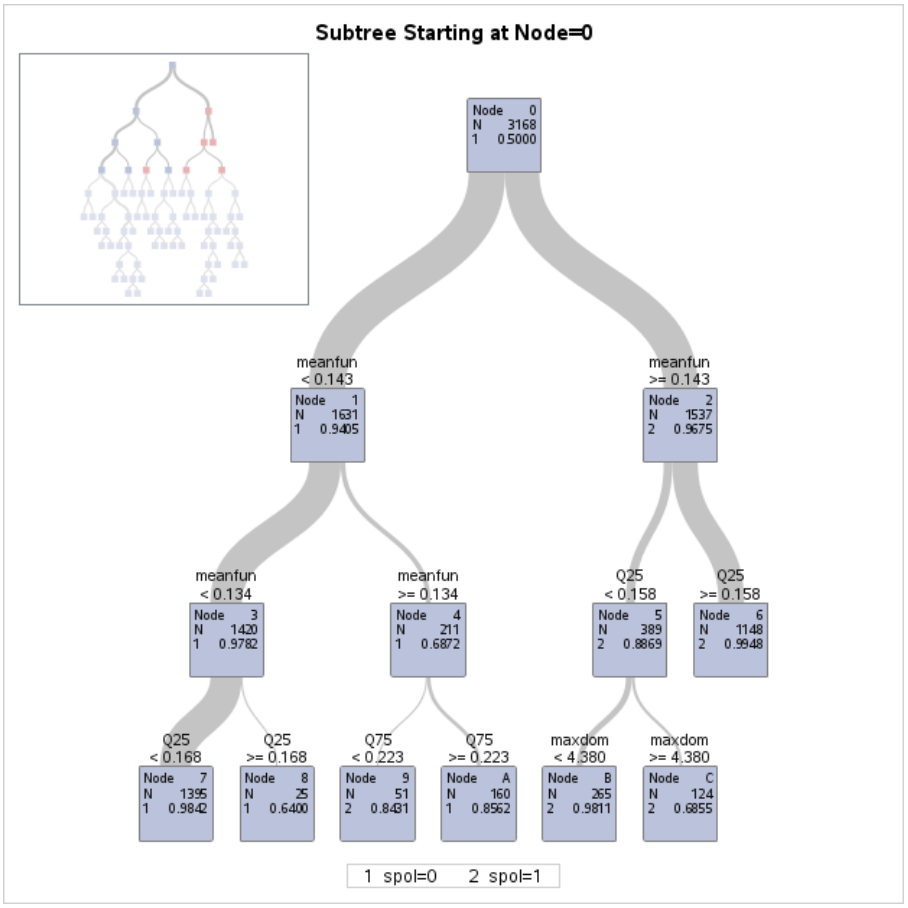
Slika 3.54: ASE kao funkcija parametara složenosti troškova

Za daljnju analizu koristit ćemo prvi kod, odnosno podstablo sa 36 listova. Na slici 3.55 imamo pregled cijelog stabla, te boja trake u svakom listu označava

najčešću vrijednost varijable *spol* (0=muškarci ili 1=žene) i predstavlja razinu klasifikacije dodijeljenju na temelju svih observaciju u tom čvoru. Visina stupca u traci pokazuje udio observacija u tom čvoru koji imaju najčešću razinu klasifikacije. Na slici 3.56 prikazan je dijagram stabla koji prikazuje više detalja o čvorovima i podijelama u prva četiri nivoa stabla. Tom slikom otkrivamo model koji se lako tumači.



Slika 3.55: Pregled stabla



Slika 3.56: Prvih četiri nivoa stabla

Prva podijela se temelji na varijabli *meanfun*. Postoji 1631 observacija čije vrijednosti varijable *meanfun* su manje od 0.143 (čvor 1), a ženski spol je prisutan na samo oko 5.95% ($(1-0.9405)\cdot 100\%$) njih. Očigledno, što je niža vrijednost srednje fundamentalne vrijednosti glasa to je manja vjerojatnost da će osoba biti ženskog spola. Daljnja podjela čvora 1 se još vrši prema varijabli *meanfun*, odnosno 1420 observacija je manje od 0.134 (čvor 3) i ženski spol je prisutan na svega 2.18%. Podjela čvora 3 se temelji na varijabli *Q25*. Imamo da ženski spol predstavlja 1.50% od 1395 observacija za koje je *meanfun* < 0.134 i *Q25* < 0.168. Analogno gledamo i za druge čvorove, na primjer za čvor B pročitamo da je ženski spol predstavlja 98.11% od 265 observacija za koje je *meanfun* ≥ 0.143, *Q25* < 0.158 i *maxdom* < 4.380. Koristeći naredbu PLOTS = ZOOMEDTREE opcija, daje nam mogućnost da prikažemo podstablo koje počinje u nekom izabranom čvoru.

Sljedeće tri tablice prikazuju procjenu točnosti odabranog klasifikacijskog sta-

bla.

Tablica 3.29: Konfuzijska matrica za klasifikaciju po *spolu* (ispis iz SAS-a)

Model-Based Confusion Matrix

Actual	*Predicted (0/1)*		*Error Rate*
0	1570	14	0.0088
1	9	1575	0.0057

Tablica 3.29 prikazuje konfuzijsku matricu iz koje možemo pročitati koliko je observacija točno odnosno netočno procijenjeno prema spolu. Muški spol je označen s 0, te iz dane tablice pročitamo da je 1570 observacija točno procijenjenih, a 14 observacija netočno. Odnosno, stopa pogreške da krivo procijenimo muški spol je 0.88%, a da krivo procijenimo ženski spol iznosi 0.57%. Model temeljen na konfuzijskom matrici se ponekad naziva i izmijenjena konfuzijska matrica (eng. *re-substitution confusion matrix*) zbog toga jer se rezultat koji se dobije na temelju trening podataka primijenjuje na testne podatke.

Tablica 3.30: Statistika prilagodbe za klasifikaciju po *spolu* (ispis iz SAS-a)

Model-Based Fit Statistics for Selected Tree

N Leaves	*ASE*	*Mis-class*	*Sensitivity*	*Specificity*	*Entropy*	*Gini*	*RSS*	*AUC*
40	0.00659	0.0073	0.9912	0.9943	0.0455	0.0132	41.7579	0.9975

U tablici 3.30 prikazane su statistički podaci za klasifikacijsko stablo. Stopa pogrešne klasifikacije za model je jako niska (0.73 %), ali odgovarajuća osjetljivost, koja prikazuje točnost predviđanja spola prema onim karakteristika koje odgovaraju određenom spolu, iznosi 99.12 %. Zadnja tri stupca u tablici 3.30 su statistike vezane uz *ROC* krivulju koja je prikazana na slici 3.57. AUC statistika je površina ispod *ROC* krivulje i u našem slučaju iznosi 0.9975 što nas upućuje kako je ovaj model praktički savršen.

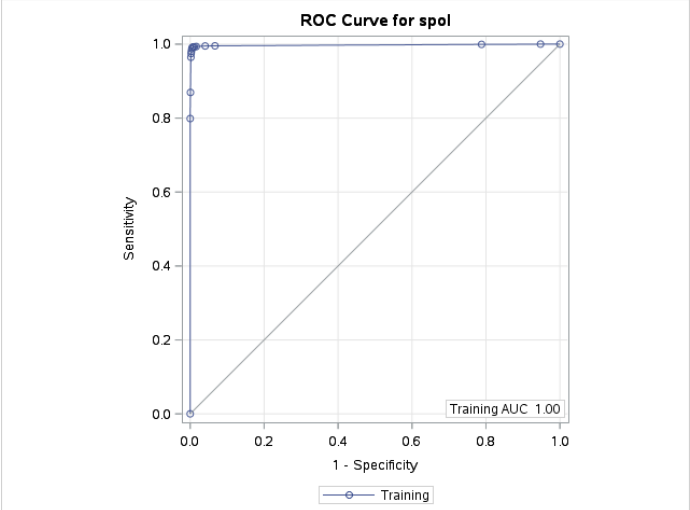
Tablica 3.31: Značajne varijable (ispis iz SAS-a)

Variable Importance

Variable	*Training (Relative / Importance)*		*Count*
meanfun	1	36.915	5
Q25	0.18	6.6462	8
Q75	0.1789	6.6027	4
maxdom	0.1502	5.544	5
minfun	0.1041	3.8444	4
meanfreq	0.103	3.8039	3
sd	0.0884	3.2632	4
sfm	0.068	2.5084	2
mindom	0.0605	2.233	1
skew	0.0558	2.0581	1
modindx	0.046	1.6971	1
maxfun	0.0447	1.6518	1

Tablica 3.31 nam je vrlo značajna jer na temelju nje vidimo koje varijable i kojim redoslijedom su statističke značajne. Primijetimo da je varijabla *meanfun* najznačajnija varijabla i kod CART analize, varijabla *Q25* sljedeća koja se pojavljuje kao druga značajna i kod univarijatne logističke regresije.

Slika 3.57 prikazuje *ROC* krivulju, koja je dobije samo u slučaju dihotomne varijable. AUC statistika i vrijednost *ROC* krivulje su izračunate iz trening podatka, u našem slučaju na temelju svih podataka. Ukoliko smo početne podatke razdvojili na dvije particije, tako da se u jednoj particiji nalaze trening podaci a u drugoj testni podaci. Tada na grafu dobijemu dodatnu *ROC* krivulju i AUC statistiku čije se vrijednosti izračunavaju iz testnih podataka.



Slika 3.57: *ROC* krivulja za klasifikaciju po *spolu*

Bibliografija

[1] *SAS OnDemand for Academics*, `https://odamid.oda.sas.com/SASODAControlCenter/`.

[2] P.D. Allison, *Logistic Regression Using SAS: Theory and Application*, SAS Institute Inc., NC, USA, 1999.

[3] David W. Hosmer i Stanley Lemeshow, *Applied Logistic Regession*, John Wiley and Sons, Inc., 1989.

[4] A.I. Izenman, *Modern Multivariate Statistical Techniques: Regression, Classification and Monidold Learning*, Springer Science + Business Media, LLC, 2008.

[5] J. Morgan, *Classification and Regression Tree Analysis*, 2014.

[6] SAS/STAT, *SAS/STAT 14.1 User's Guide: The HPSPLIT Procedure*, 2015.

[7] M. Huzak, *Statistički praktikum 1*, 2015.

[8] A. Jazbec, *Odabrane statističke metode u biomedicini*, 2016.

Sažetak

U ovom radu objašnjeni su glavni koncepti logističke regresije i klasifikacijskog stabla. Nakon kratkog uvoda i teorijske podloge, koristeći logističku regresiju i klasifikacijsko stablo, izvršili smo statističko predviđanje spola govornika prema akustičnim karakteristikama glasa i govora. Statistička analiza podataka izrađena je u statističkom programskom paketu SAS Studio ([1]).

U primjeni je često varijabla od interesa diskretna, odnosno varijabla koja poprima vrijednosti iz prebrojivog skupa mogućih vrijednosti. Model logističke regresije se najčešće koristi za analizu upravo takvih podataka. Logistička regresija se koristi u različitim područjima, uključujući strojno učenje, najčešće u medicini i društvenim znanostima.

Stabla odlučivanja spadaju u pristupe prediktivnih modela koji se koriste u statistici, u rudarenju podataka i strojnom učenju, te je njihova prednost jednostavnost i razumljivost metode. Metoda klasifikacijskih stabala temelji se na binarnom stablu, odnosno ukoliko je ciljna značajka diskretan skup vrijednosti.

Obje metode pokazale su se uspješnima u predviđanju spola govornika.

Summary

This thesis explains the main concepts of logistic regression and classification trees and demonstrates their application. After a brief introduction and a theoretical background, using logistic regression and classification tree, an estimation of the speaker's gender based on the acoustic characteristics of their voice and speech was performed. The statistical analysis of the data was made in the statistical program package SAS Studio ([1]).

In applications the variable of interest is often discrete, that is a variable that receives values from a countable set of possible values. The logistic regression is most often used to analyze such data. Logistic regression is used in various fields, including machine learning, many medical fields, and social sciences.

Decisions trees belong to the predictive modeling approaches used in statistics, *data mining* and machine learning. One of their main advantages is simplicity. The method of classification trees is based on a binary tree, i.e. if the target feature is a discrete set of values.

Both methods have proven to be successful in estimating the speaker's gender.

Životopis

Rođena sam 30. rujna 1992. godine u Zagrebu. Nakon završene osnovne škole Matije Gubec, 2007. godine u Zagrebu upisujem XV. gimnaziju (MIOC) informatički smjer. 2011. godine upisujem Preddiplomski sveučilišni studij Matematika na Matematičkom odsjeku Prirodoslovno-matematičkog fakulteta u Zagrebu. Nakon završenog preddiplomskog studija, 2015. godine upisujem Diplomski sveučilišni studij Matematička statistika na istom fakultetu.

Od početka studiranja dajem instrukcije učenicima osnovnih i srednjih škola, te pripreme za maturu. Prva četiri ljeta studiranja sam radila kao zamjena glavne tajnice u EurocomputerSystems-u (ECS). Od siječnja do travnja 2016. godine bila sam na praksi u PwC-u (PricewaterhouseCoopers) u RAS (Risk Assurance Service) odjelu, te krajem kolovoza 2017. godine počinjem raditi u PwC-u.